DOCUMENT RESUME

ED 301 587 TM 012 483

AUTHOR Archbald, Doug A.; Newmann, Fred M.
TITLE Beyond Standardized Testing: Assessing Authentic Academic Achievement in the Secondary School.
INSTITUTION National Association of Secondary School Principals, Reston, Va.; National Center on Effective Secondary Schools, Madison, WI.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE 88
GRANT OERI-G-008690007
NOTE 74p.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Books (010)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Academic Achievement; *Achievement Tests; *Educational Assessment; Educational Quality; *High School Students; Organizational Effectiveness; Secondary Education; *Standardized Tests; *Student Evaluation
IDENTIFIERS Authenticity

ABSTRACT

This book was designed as an assessment of standardized testing and its alternatives at the secondary school level. More specifically, a framework for thinking systematically and creatively about assessment, a review of the uses and limitations of standardized tests of general achievement, and descriptions of several methods that may offer more helpful approaches to assessment are provided. All three specific components are grounded in a broad perspective that calls attention to the purposes of assessment, levels of assessment, and two critical issues (authenticity and multiple indicators). The nature of authentic academic achievement is discussed, and approaches to assessing authentic academic achievement are forwarded. Assessment of organizational academic quality is addressed, and implementation of assessment programs is reviewed. A discussion on the uses and limitations of standardized tests is appended. A 66-item list of references is included. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED301587

Beyond Standardized Testing

Assessing Authentic Academic Achievement In the Secondary School

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Doug A. Archbald
Fred M. Newmann

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

T. KOEFNER

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

National Association of Secondary School Principals

TM012483

TM

Beyond Standardized Testing

Assessing Authentic Academic Achievement In the Secondary School

Doug A. Archbald
Fred M. Newmann

About the Authors:

Doug A. Archbald is assistant researcher, Center for Policy Research in Education at the University of Wisconsin-Madison.

Fred M. Newmann is professor of curriculum and instruction and director of the National Center on Effective Secondary Schools at the University of Wisconsin-Madison.

Copyright 1988
National Association of Secondary School Principals
1904 Association Dr., Reston, Va. 22091
(703) 860-0200
All rights reserved

Executive Director: Scott D. Thomson
Director of Publications: Thomas F. Koerner
Project Editor: Patricia Lucas George
Technical Editor: Eugenia Cooper Potter

This paper was prepared at the National Center on Effective Secondary Schools, School of Education, University of Wisconsin-Madison which is supported in part by a grant from the Office of Educational Research and Improvement (Grant No. G-008690007). Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of this agency or the U.S. Department of Education.

Prepared in cooperation with the National Association of Secondary School Principals.

Contents

Introduction v

1. What Is Authentic Academic Achievement? 1

2. Approaches to Assessing Authentic Academic Achievement 5

3. Assessing Organizational Academic Quality 34

4. Implementing Assessment Programs 45

Appendix: Uses and Limitations of Standardized Tests 52

References 65

ACKNOWLEDGMENTS

For thoughtful reviews of this manuscript we are indebted to Richard DuFour (Principal, Adlai Stevenson High School, Prairie View, Ill.); Joseph Levanto (Principal and Superintendent, Norwich Free Academy, Norwich, Conn.); Glynn Ligon (Executive Director, Department of Management Information, Austin Independent School District, Austin, Tex.); Arthur Powell (Director, Commission on Educational Issues, National Association of Independent Schools, Boston, Mass.); Diana Schmelzer (Principal, South Lakes High School, Reston, Va.); Richard Sagor (Assistant Superintendent, West Lynn School District, West Lynn, Oreg.); and Grant Wiggins (Director of Research, Coalition of Essential Schools, Brown University, Providence, R.I.).

Sincere thanks also to representatives from the following schools and agencies for permission to describe their work (in the order of their appearance in the text): Milwaukee Public Schools, Milwaukee, Wis.; School District No. 12, Adams Country, Northglenn, Colo.; The Assessment of Performance Unit, Department of Education and Science, Welsh Office, Northern Ireland; The National Assessment of Educational Progress, Princeton, N.J.; Alverno, College, Milwaukee, Wis.; Frontenac Secondary School, Kingston, Ontario; Jefferson County Open High School, Evergreen, Colo.; Walden III High School School, Racine, Wis.; Learning Unlimited North Central High School, Indianapolis, Ind.; Adlai E. Stevenson High School, Prairie View, Ill.; Saint Paul Public Schools, St. Paul, Minn.; The California State Department of Education.

Introduction

THE EDUCATION REFORM MOVEMENT of the 1980s has cast a critical spotlight on high schools. Policymakers, business leaders, school boards, and parents want accountability, proof that their investment in education produces higher levels of achievement for all students.

This mounting pressure has led to increased reliance on testing to monitor achievement, especially on competency tests and norm-referenced standardized tests developed by authorities beyond the classroom. At the same time, a number of authorities, from teachers to policymakers, have called for alternatives to standardized testing that might offer more informative and authentic indicators of the kinds of achievement schools ought to promote.

This book has three main purposes:

- To offer a framework for thinking systematically and creatively about assessment
- To review the uses and limitations of standardized tests of general achievement
- To describe a variety of methods that may offer more helpful approaches to assessment. All three are grounded in a broad perspective that calls attention to purposes of assessment, levels of assessment, and two critical issues: authenticity and multiple indicators.

Why Assessment?

Educational assessment usually serves three goals:

- To show the extent that schools, students, or teachers have met their objectives
- To tell them what they might do to improve
- To select the most promising students for college, the most effective teachers or schools for recognition, or the most deserving students or schools for financial assistance.

It is important to distinguish among the purposes of educational assessment. The collection of information for one purpose will not necessarily serve another. For example, standardized tests of general achievement such as the SATs were designed to give information about students most likely to succeed

in college, not to indicate the effect of school on student achievement or what students, teachers, or schools might do to improve.

Levels of Assessment

Individuals, groups, or organizational units such as schools, districts, or states can be assessed. Students, parents, and school personnel must learn what the individual student has achieved. It is also important to assess the overall impact that the school has on large groups of students. When and how should measures of individual achievement be aggregated to the school level, and can they provide meaningful indicators of school productivity?

Individual student test scores may be useful for describing a student's achievement relative to others, but aggregating these scores to create school means will produce weak information about school quality. Such scores alone fail to detect the possibility of disparate outcomes for different groups of students within the school.

An assessment design, therefore, must be constructed with an awareness not only of the general purpose(s) to be served, but also of the ways in which the information might be used to reach conclusions about both individuals and groups or organizations.

Two Critical Issues

A fundamental issue lies beneath all testing exercises: does the information collected represent an accurate estimate of worthwhile knowledge and mastery? Among other criticisms, traditional tests have been faulted for neglecting the kinds of competence expressed in authentic, "real life" situations beyond school—speaking, writing, reading, and solving mechanical, biological, or civic problems.

The concern that assessment should measure not just any kind of achievement, but rather, valuable or meaningful forms of mastery, can be summarized as the problem of authenticity. Judging the authenticity of achievement is a complex problem, but it must be faced. Explicit criteria for authenticity should guide examination of both traditional and alternative forms of testing.

Another critical issue is when to use multiple indicators of achievement. Any single indicator, such as a percentile rank or a dropout rate, limits the amount of information conveyed and is subject to error. In purchasing a car, most of us are interested not only in price or gas mileage, but also in safety record, reliability and service record, resale value, and warranty. Teachers, after all, assign grades, based not on one test, but on several types of student performance (homework, class participation, special projects, quizzes, major tests).

At times, of course, single indicators may be important, such as when a school works to increase attendance or when a student devotes special effort to improve a particular skill. In judging overall performance, however, multiple indicators are usually needed. Unfortunately, the desire for *simple* numbers has created accountability pressures that can undermine the utility, validity, and fairness of both individual and organizational assessment.

There are no foolproof approaches to high school assessment. Improvements will demand serious effort by principals and others. Ultimately, improved assessment should bring clarity and consensus on educational purpose, help for teachers to be more effective, and increased student commitment to academic learning.

The book begins by proposing criteria to define authentic assessments of academic achievement (Chapter 1). Next it presents alternatives likely to yield authentic measures of individual achievement (Chapter 2). Multiple schoolwide indicators of academic success that include more than exit measures of individual student performance are discussed in Chapter 3. Finally, Chapter 4 presents a proposal for implementing these ideas in a high school.

This book was developed to respond to the widely held concern that standardized norm-referenced tests of general achievement offer inadequate indicators of students' academic accomplishments. An up-to-date review of the uses and limitations of standardized tests is provided in the Appendix.

CHAPTER 1

What Is Authentic Academic Achievement?

ASSESSMENT IN EDUCATION is not a value-neutral enterprise. It reflects value commitments and ideology, though these are not always apparent. We endorse some outcomes, such as high test scores, without scrutinizing the tasks on which they are based, and sometimes we allow the technology of testing to shape the curriculum. We often assume that educational purposes are not controversial, yet in an increasingly pluralistic and changing society, schools have to work more self-consciously to build consensus on educational standards.

In coping with controversy about appropriate educational standards, we must be careful to see that assessment exercises are driven by the human purposes we want education to serve. Once these basic goals are formulated, indicators should be devised to estimate how many of and how well these accomplishments have been mastered. We have typically relied on school grades, credits, test scores, degrees, and honors as indicators of mastery. As these factors become increasingly significant in students' lives, we must face at least two problems related to the question of educational purpose.

First, most traditional assessment indicators communicate very little about the quality or substance of students' specific accomplishments. For example, even though we know a graduate's grade point average or a school's percentage of failing grades, the level of mastery is unclear unless we also know the nature of tasks and tests used in the school to measure that mastery.

Second, the type of learning actually measured is often considered trivial, meaningless, and contrived by students and adult authorities. A valid assessment system provides information about the particular tasks on which students succeed or fail, but more important, it also presents tasks that are worthwhile, significant, and meaningful—in short, *authentic*.

Assessments that provide little substantive information and lack authenticity undermine the legitimacy not only of the numerical indicators, but of the educational enterprise itself. This can depress student learning, teacher commitment, and public support. In contrast, informative assessment of authentic achievement can invigorate teaching, learning, and public support. Such assessments should consider tasks that meet at least three criteria: disciplined inquiry, integration of knowledge, and value beyond evaluation.

Disciplined Inquiry

Authentic academic achievement reflects the kinds of mastery demonstrated by experts who create new knowledge. Scientists, historians, jurists, literary and artistic critics, philosophers, mathematicians, linguists, sociologists, journalists, and other experts share a common approach to work—disciplined inquiry.

Disciplined inquiry consists of three features. First, it depends on prior substantive and procedural knowledge considered essential to understanding problems within a field. For new knowledge to be considered valid, it must respond in some way to the public knowledge base, even if it rejects aspects of prior knowledge.

Second, disciplined inquiry tries to develop in-depth understanding of a problem rather than passing familiarity with or exposure to pieces of knowledge. Prior knowledge is mastered, not to become literate about a broad survey of topics, but to facilitate complex understanding of relatively limited, special problems.

Finally, the ultimate point of disciplined inquiry is to move beyond knowledge that has been produced by others; that is, to assemble and interpret information, to formulate ideas, to make critiques that cannot be easily retrieved from the existing knowledge base.

Most adolescents, of course, cannot be expected to achieve the level of disciplined inquiry demonstrated by experts. However, the authenticity of students' academic achievement will be enhanced if the tasks on which they are evaluated approximate these attributes of disciplined inquiry.

Achievement in science, for example, could place more emphasis on the development, execution, and reporting of a single experiment. In history and social studies, intensive research using primary source materials could help students evaluate generalizations stated in their textbooks. In studying a literary work, one might aim toward students' clarifying and defending their own views of alternative interpretations.

Integration of Knowledge

To understand scientific theories, literary and artistic masterpieces, architectural and mechanical designs, musical compositions, or philosophical ar-

guments, we must ultimately consider them as wholes, not as collections of knowledge fragments. For example, teaching students the separate roles of each character does not provide them with an authentic understanding of a play. The characters must be studied in the context of the overall pattern of plot, literary technique, historical context, and philosophical position. Similarly, an authentic understanding of a molecule or atom should integrate the "parts" into broader conceptions of matter or energy.

Assessments should, therefore, tap the student's grasp of relationships. Too often tests of achievement ask the student only to show comprehension of unrelated knowledge fragments: definitions of terms; short descriptive identifications of people, things, events; or numerical solutions to problems. Students demonstrate proficiency by giving short responses, as in a television quiz show, where answers bear little relation to one another. In such a case, knowing the correct answer may contribute to more integrated understanding of the topic, but cannot be considered an indicator of it.

Authentic academic achievement should integrate knowledge in two ways. Not only must students be challenged to understand integrated forms of knowledge, they must also be involved in the production, not simply the reproduction, of new knowledge, because this requires knowledge integration.

Value Beyond Evaluation

Demonstrations of disciplined inquiry are most meaningful when achievement has aesthetic or utilitarian value apart from determining the competence of the learner. When people write letters, news articles, insurance claims, poems; when they speak a foreign language; when they develop blueprints; when they create a painting, a piece of music, or build a stereo cabinet, they demonstrate achievements that have a special value missing in tasks contrived only for the purpose of assessing knowledge (such as spelling quizzes, laboratory exercises, or typical final exams).

What characterizes these tasks? Authentic demonstrations of mastery often share three features uncommon in most school testing situations: the production of discourse, things, or performances; flexible use of time; and collaboration with others.

- *Production of discourse, things, performances.* Beyond school we demonstrate knowledge by providing original conversation and writing, by repairing and building physical objects, and by producing artistic, musical, and athletic performances.

  In contrast, assessment in school usually asks students to identify the discourse, things, and performances that others have produced (for example, by recognizing the difference between verbs and nouns, between socialism and capitalism; by matching authors with their works; by correctly labeling rocks and body parts).

- *Flexible use of time*. The significant achievements of disciplined inquiry often cannot be produced within rigidly specified time periods. Adults working to solve complicated problems, to compose effective discourse, or to design products are rarely forced to work within the rigid time constraints imposed on students such as the 50-minute class or the two-hour examination period.

  Standard, predetermined time schedules based on bureaucratic procedures for managing masses of students and diverse course offerings, rather than on the time requirements of disciplined inquiry, can reduce the authenticity of student achievement. Achievements in non-instructional settings (such as journalistic writing, financial analysis, counseling, cooking) do, of course, involve deadlines and time limits, but because these tend to be determined more by the nature of the work than by the requirements of institutional management, they are less likely to diminish the authenticity of achievement.
- *Collaboration*. Achievements outside school often depend on the opportunity to ask questions of, to receive feedback from, and to count on the help of others, including peers and authorities. In contrast, typical assessment of school achievement focuses primarily on what the student can accomplish while working alone. Assessment tasks that deny opportunities to cooperate can thereby diminish the authenticity of the achievement.

In short, authenticity will be enhanced if assessment tasks have aesthetic or utilitarian value beyond instructional evaluation, and this is most likely when the tasks involve student production, rather than reproduction, of knowledge, flexible use of time, and access to help from others in completing the task.

Summary

Improved assessment of academic achievement in high schools will require more than better techniques such as advanced computer services. To ensure that assessment tasks, whether teacher-produced exercises or externally developed tests, provide information about authentic forms of achievement, we emphasize three criteria: disciplined inquiry, integration of knowledge, and value beyond evaluation. Staff members must develop clear standards for the more specific content, skills, and attitudes that shape the substance of assessment tasks, but these criteria offer a foundation for evaluating existing assessment exercises and the development of new ones. Present approaches to assessment often fall considerably short of these criteria, and workable alternatives may be hard to find. It will take commitment, patience, and creativity to develop and adapt new approaches to the real world of high schools. Eventually this may require fundamental changes in the nature of schooling itself. We can move incrementally, however, by building exercises that respond to some, if not all, of the criteria and by using these for some, if not all, the indicators of student achievement.

CHAPTER 2

Approaches to Assessing Authentic Academic Achievement

THIS CHAPTER DESCRIBES assessment practices that tend to meet one or more of the criteria of authentic achievement presented in Chapter 1. The focus here is on assessment of individual students rather than the school as a whole. The examples come from high schools in four countries and represent a broad spectrum of approaches, from scored writing samples to public exhibitions of diverse competencies.[1]

The chapter is divided into three parts: discrete competency tasks; exhibitions; and portfolios and profiles. Most of the examples in Part One describe assessment of discrete academic competencies under standard conditions. These techniques yield quantitative information that can be used to describe progress or change in individual students as well as comparative success of groups of students. The examples assess such competencies as writing, speaking and listening, analyzing, and problem solving.

The exhibitions described in Part Two are public demonstrations of academic mastery. The exhibition is meant to reflect competence on challenging tasks that do not have a single clear solution and that require analysis, creativity, and usually considerable integration of knowledge. Exhibitions result in products useful and interesting in their own right; such as presentation of a research report or literary essay, a music recital or art exhibition, or the demonstration of a mechanical invention.

[1]The examples were selected through a search of the literature, an announcement in the NASSP *NewsLeader* (October 1986), and solicitation of suggestions from authorities in numerous research centers and professional organizations.

Portfolios and profiles described in Part Three provide summaries or inventories of many facets of individual accomplishment.

People who use these approaches understand the importance of assessment as a learning and management tool. As stressed in Chapter 1, assessment drives the curriculum; it signals what counts. When we test for trivial or inauthentic achievement, teaching and learning are corrupted and "teaching to the test" becomes a dirty word. But if we test for authentic forms of achievement, teaching to the test is appropriate and desirable.

Course syllabi might well *begin* by presenting the "final exam" or other assessment tasks that require disciplined inquiry, integration of knowledge, and the production of knowledge that has value beyond evaluation. These tasks should cast a shadow back on the curriculum, serving as a constant target for study and practice, as when artists prepare for performances, when carpenters build a house, when athletes prepare for the big game.[2] In short, tests, projects, and performances that demonstrate authentic academic achievement are valuable not only as assessment devices, but as guides to focus and to inspire teaching and learning.

Part One: Tests of Discrete Competencies

To measure proficiencies under standard conditions that permit comparisons over time and between student groups, and to assign numbers that stand for varying degrees of success so that assessment can be summarized in a simple indicator and aggregated, it is often useful to break achievement into discrete parts. The examples that follow show forms of competency testing that produce such indicators and that also meet several criteria for authenticity in the areas of language performance and analytical problem solving. These procedures can also be used to assess mastery of specific curriculum content.

Language Performance

Questions in the following approaches are designed to assess verbal competence apart from grasp of particular subject matter.

---

[2]The American College Testing corporation, through the College Outcome Measures Project (Forrest and Steele, 1982), provides a useful conceptual framework for thinking about educational outcomes. It has developed a set of instruments, available from ACT, for the assessment of proficiencies in applying specific facts and concepts in work, family, and community roles.

1. Writing Assessment: Holistic Scoring

Every spring, each ninth grader in the Milwaukee, Wis., public schools writes an essay and a business letter, which are graded using holistic scoring procedures. The assessment identifies students who need extra writing instruction prior to graduation. It also helps to reveal strengths and weaknesses in writing instruction among the district's ninth graders.

At a designated time and day, all ninth graders are given the same set of instructions and 90 minutes to write an essay and a business letter. The majority of the students finish in this time, but an extra half hour is given to those who need additional time.

The essay, which counts twice as much as the business letter, stresses writing as expression, more than as an instrument for practical ends. Students are required to produce ideas and to support opinions about familiar topics. Recently students were asked to write about problems with their friends, neighborhoods, or schools, and present and support three ideas for improving their school. Examples of the business letter include a job application letter and a consumer complaint letter.

Scoring

English teachers score the writing during a week in June. They first develop consensus on standards. A small team of experienced readers selects about 35 papers (for each writing task) from the pool of student papers to represent the range of quality of student writing. These are called "benchmark" papers. The benchmark papers are duplicated and a set is given to each of the approximately 60 readers.

The business letters are scored first, because they are easier to grade. Guidelines are handed out with written criteria corresponding to each of four possible scores, from 1 (highly flawed, not competent) to 4 (competent, clear mastery). Specific criteria for each of the scores are indicated in Figure 2.1.

Then, using the benchmark papers, the group grades single papers to reach consensus on standards. When the group achieves consistency in its scoring standards with the benchmark papers, the scoring process is ready to begin.

Packets of 20 papers are given to each reader (student anonymity preserved) for scoring. After all the papers have been scored once, they are read and scored again by a second reader. The first readers use codes to disguise the scores they assign to papers so that the second readers are not influenced by the first rating. Also, papers are shuffled between packets to vary the assignment of papers to readers. The scores of the two readers are summed to produce the final score for each paper.

A leader and a few master readers coordinate the scoring process to ensure that the packets are exchanged properly; conduct intermittent "benchmarking" sessions to ensure that the standards remain consistent; help readers pace themselves; answer questions during the scoring process; and re-read and make

Figure 2.1

Criteria for Holistic Writing Assessment

1—Highly flawed—Not competent
- Ideas poorly communicated
- Frequent usage errors (such as: agreement, pronoun misuse, tense)
- Incorrect or erratic use of capitalization, punctuation, and spelling conventions
- Sentence fragments and run-ons; few complete sentences
- No concept of paragraph construction

2—Unacceptable—Not competent
- Poor organization of ideas
- Frequent usage errors (such as: agreement, pronoun misuse, tense)
- Inconsistent use of capitalization, punctuation, and spelling conventions
- Sentence fragments and run-ons; few complete sentences
- Poor topic sentence; flawed paragraph development

3—Minimally competent—Acceptable
- Ideas sufficiently organized and communicated
- Only occasional usage errors (such as: agreement, pronoun misuse, tense)
- Basically correct capitalization, punctuation, and spelling
- Minimal number of sentence errors (fragments or run-ons)
- Paragraphs have topic sentences, supporting ideas, closing sentences
- Some attempt at paragraph transition

4—Competent—Clear mastery
- Ideas clearly communicated and of a fairly mature quality
- No usage errors
- Correct capitalization, punctuation, and spelling
- No fragments or run-ons
- Paragraphs have topic sentences, supporting ideas, closing sentences, and are developed in a mature fashion
- Excellent vocabulary
- Effective paragraph transitions

0—Represents a paper that is illegible or off the point.
A non-response is also a 0 paper.

Source: Division of Curriculum and Instruction, Department of Elementary and Secondary Education, Milwaukee (Wis.) Public Schools.

a final judgment on all papers with scores that differ by more than one point. In Milwaukee's scoring sessions, about 1 percent of papers must be read a third time.[3]

After the business writing samples are scored, the entire process, including the steps for setting standards, is repeated for the essays. A student's final score for the overall writing task is weighted to give the essay twice the value of the business letter.

At a debriefing session, the readers share observations about spelling, punctuation, grammar, organization, and the content of the papers. This provides feedback to teachers and central office coordinators about elements of writing as well as insights into students' concerns and values. Teachers have described the process as "professionalizing"—an opportunity to seriously "talk shop" and to reflect on educational purpose, standards, and evaluation in writing.

The writing test also legitimates teachers' efforts to teach writing as a process in which organization and revision of written ideas are pivotal.[4] Central office administrators favor the benefits of staff development and the overall contribution to the clarification of standards in the district's writing program.

2. Writing Assessment: Analytical Scoring

Adams County School District #12 in Northglenn, Colo., assesses writing competence at each grade each year using primary trait assessment.

All teachers in the district receive a packet describing the purpose of the test, the days on which the test is to be administered, and how to administer it. Students are provided with lined writing sheets, written instructions, and the question to which they must write a response.

Students complete the writing task in two 45-minute class periods. The first period is devoted to outlining and writing a first draft. The next day, students write and turn in the final product.

Paid readers, English teachers from the district and competent readers from the community, grade the writing samples. Each writing sample has two readers. Reliability checks similar to those described in the holistic scoring process for Milwaukee are used.

---

[3]According to research on the method, the inter-rater reliability of holistic grading is high, in the .7 to .9 range (Hogan and Mishler, 1981). The extent to which inter-rater scores match depends on the level of training of the raters, the clarity of guidelines, and the range in scores used in judging. Generally, fewer than 10 percent of papers need to be re-read; 5 percent or less is considered desirable. For discussions of specific programs and issues in the assessment of writing, see Greenberg, Wiener, and Donovan, 1986; and articles in the Spring 1984 issue of *Educational Measurement: Issues and Practice*.

[4]For further discussion see Greenberg, Wiener, and Donovan, 1986.

Readers score the writing samples according to multiple criteria (as opposed to a single holistic judgment of quality). Each student writing sample is rated 1 to 5 on each of the following criteria:

- *Organization*—measures the students' ability to logically organize ideas into paragraphs that develop their ideas and to combine paragraphs into well-sequenced essays
- *Sentence Structure*—measures students' ability to write complete sentences and to vary syntax
- *Usage*—measures students' ability to select the correct words to carefully communicate a precise message
- *Mechanics*—measures students' ability to capitalize, punctuate, and spell correctly
- *Format*—measures students' ability to form letters and numbers correctly and to use correct margins and letter format when applicable.

Figure 2.2 shows the performance standards and writing criteria for the eleventh grade writing sample. The numbers across the top of the chart represent the 1 to 5 rating range. The numbers in the righthand column show the weighting scheme. In computing scores, format is the least important trait. Mechanics counts four times as much as format; organization, six times as much as format.

When the scoring is completed, the results are processed by computer. The primary purpose of the writing assessment is for district-level monitoring, but teachers also use the results for diagnostic and placement purposes. Student, classroom, and school results can be compared to district norms using the total score and the individual trait scores.

3. The Assessment of Speech

In 1977, the Assessment of Performance Unit (APU) was established in the Department of Education and Science in Northern Ireland.[5] Each year, using a variety of measures, the language performance of a large national sample of students was assessed. Initially, reading and writing were assessed on a yearly basis. In 1982, the assessment of oracy, or speaking and listening, began.

The APU's main purpose is to monitor student language performance at a national level, with a rationale similar to that of the National Assessment of Educational Progress in the United States. Recently the APU published a handbook for practitioners describing the methods of the oracy assessment program. The APU also encourages schools to develop their own programs to assess the speaking and listening competencies of students. This project illustrates the importance of oral discourse and provides some guidelines for assessing it.

---

[5]For more information about the APU and a complete description of the assessment tasks described here, see: Gorman, T., White, J., and Brooks, G. (1982); Gorman, T. (1986); and Maclure, M. and Hargreaves, M. (1986).

Figure 2.2

Criteria for Analytical Scoring

	1	2	3	4	5	
Organization	Little or nothing is written. The essay is disorganized, incoherent, and poorly developed. The essay does not stay on the topic.		The essay is not complete. It lacks an introduction, well-developed body or conclusion. The coherence and sequence are attempted, but not adequate.		The essay is well-organized. It contains an introductory supporting and concluding paragraph. The essay is coherent, ordered logically, and fully developed.	x6
Sent. Str.	The student writes frequent run-ons or fragments.		The student makes occasional errors in sentence structure. Little variety in sentence length or structure exists.		The sentences are complete and varied in length and structure.	x5
Usage	The student makes frequent errors in word choice and agreement.		The student makes occasional errors in word choice or agreement.		The usage is correct. Word choice is appropriate.	x4
Mechanics	The student makes frequent errors in spelling, punctuation, and capitalization.		The student makes an occasional error in mechanics.		The spelling, capitalization, and punctuation are correct.	x4
Format	The format is sloppy. There are no margins or indentations. Handwriting is inconsistent.		The handwriting, margins, and indentations have occasional inconsistencies—no title or inappropriate title.		The format is correct. The title is appropriate. The handwriting, margins, and indentations are consistent.	x1

Source: Adams County School District #12, 11285 Highline Dr., Northglenn, Colo. 80203

Several principles guide the APU's assessment of oracy:

1. Assessment should reflect the various communicative purposes of oral communication.
2. Oral communication is relevant across the curriculum.
3. The oral and written modes should be seen as reciprocal and integrated aspects of students' overall communicative ability.
4. Listening and speaking should be considered reciprocal and integrated aspects of students' oral communicative ability.
5. Spoken language is sensitive to context, and so assessment must consider contextual factors.

The tasks used to assess students' oral skills are categorized under five general communicative purposes:

1. Instructing/directing
2. Giving and interpreting information
3. Narrating
4. Describing and specifying
5. Discussing.

Each of these uses is broken down into a number of more specific purposes that lead to one or two specific assessment tasks.

Specific assessment tasks are designed to be as realistic as possible to encourage students to use the kind of language they would use outside the test situation; to put pupils at ease to encourage spontaneous and un-self-conscious speech; and to be stimulating and fun for the participants.

In each task, a student is asked to talk with another person or a small group of people to achieve a particular purpose. In most cases the other person is a friend the student selects; sometimes small groups consisting of the friend and other students are the listeners; in some cases the assessor is the listener.

The requirement for a realistic and relaxed setting for the students is balanced against the need for standardization, which entails a pre-scripted instructional protocol that has been carefully developed and rehearsed to sound natural and non-threatening.

Each student's performance is scored in three ways. During the oracy task, students are given a holistic score (1 - 7), as well as an "orientation to the listener" (eye contact, non-verbal gestures) score (1-5). The assessor tapes each student's oracy task, which is also evaluated later by a pair of trained assessors using analytical scoring techniques. For the analytical scoring, the categories are:

1. Sequential structure (organization) [1 - 5]
2. Lexico-grammatical features (lexical selection and syntax) [1 - 5]
3. Performance features (hesitancy/fluency, tempo/pacing, and verbal assertiveness) [1 - 3].

Following are some examples of tasks the APU developed to assess oracy.

1. *Bridges: Distinguishing among complex visual patterns*

One student (the listener) has a sheet of paper showing pictures of six different bridges. Another student (the describer) has a sheet with only two of the pictures. The describer and listener face each other so that neither can see the other's sheet.

The describer, after thinking about the pictures for a minute, is instructed to describe the bridges one at a time, and in as much detail as possible. The describer is asked to begin the description of the second bridge when he or she thinks enough information about the first bridge has been given so that the listener can correctly identify the first bridge from the six on the sheet. The listener is told not to verbally identify any bridges until both descriptions are completed. The listener cannot ask questions. This encourages the describers to provide as much detail as possible.

2. *Spider Web: Interpreting a series of events depicted in a set of drawings*

Students listen to a tape that describes how a garden spider builds its web. As they listen they examine and arrange sequentially, in accordance with the recorded description, six picture cards that illustrate different stages in the process. Finally, they must recount the stages in the process to their partner students, using the diagrams as an aid.

3. *Language and the Brain: Summarizing a short recorded message that interprets a diagram*

Students are given a diagram of the human brain and told to take notes while they listen to a tape about "language and the human brain." Afterward, they must explain and summarize the contents of the tape to listeners who have not heard the tape.

In each of these tasks, only the speaker is assessed. The listeners' responses are not a criterion for success, as the main purpose is to assess the speakers' ability to translate visual observations and oral messages into their own language, demonstrating effective orientation to the listener, good organization, proper lexico-grammatical features, and aspects of speech performance such as fluency and proper timing.

The APU has developed other tasks to assess oracy in different contexts and for different purposes. In some APU tasks, students tell the assessor information that is unique to them. For example, after being asked to describe "something you have learned recently" or "a place you know about," students respond to conversation-style questions on the subject they have chosen. Students are urged to select topics the assessor is unlikely to know much about, which places the student in the position of being a knowledgable authority.

Although oral discourse is rarely assessed systematically, a strong case can be made for its importance as a goal of schooling (Newmann, 1988). The APU assessment methods, developed from substantial research and classroom experience, illustrate how this goal might be approached.

Analysis and Problem Solving

1. Essay and Oral Exams

The essays and oral examinations required in graduate programs are often considered the most rigorous and valid tests of academic competence. They are recognized internationally for the depth with which they assess mastery of specific subjects as well as analytical skills. Usually they require the student to integrate knowledge and speak extensively. These exams are not standardized according to the tasks presented, testing conditions, and/or criteria for success, but within specific subject fields there is probably substantial consensus on the hallmarks of competent and distinguished performance.

Why aren't essay and oral exams used more frequently in high schools? Because of time constraints, students often have few opportunities to write or speak more than a sentence or two. This is unfortunate, since most people would probably agree that the best way to determine whether a person understands a subject or problem is simply to ask him or her to *explain* and to respond to questions that the explanation itself is likely to provoke. Large teaching loads and the technology of testing have perhaps obscured this important principle, but several leading educators continue to emphasize the use of written and oral language as the coin of academic mastery (e.g., Adler, 1982; Boyer, 1983; Sizer, 1984).

Figure 2.3

Mary is going to hike into a lake in the Oregon Cascades. Fifteen years earlier, Mary had been to the same lake to conduct a study for the Oregon Fish and Wildlife Commission. At that time the lake was a typical high mountain lake surrounded by coniferous trees on three sides and some alders, birches, and maples on the more level, meadow side of the lake. The lake had been a favorite fishing spot for her father and grandfather, producing many shrimp-fed rainbow trout. She learned that crawfish from the lake were good bait for the fish.

Describe the changes you think might have occurred in the plants and animals of this environment if acid rain had significantly affected the area. Make specific references to the assigned reading to support your hypotheses.

The following criteria will be used to evaluate your response:

1 point will be given if your response is clear and well organized.

1 point will be given if the response is logically supported by specific references to the background reading.

2 points will be given if the response shows "in-depth" thought, that is, a careful and thorough consideration of the possible effects that acid rain might have had on this environment.

Source: Fielding and Fiasca, in press.

Students infrequently write essays that are more than a paragraph long (Applebee, 1981; 1984), but teachers have devised many interesting assignments that ask students to demonstrate mastery by constructing original explanations and arguments. Figure 2.3 is an example of an essay question that could follow a science or social studies unit dealing with pollution.

Oral examinations in high school are rare. Students may participate in small group discussions, and occasionally even in short Socratic discussions, but these are usually not required as significant assessment exercises. To complement the kinds of oracy exercises described above, oral examinations can focus on subject matter content. These exams can take several forms, such as student debates, teacher questioning of small groups of students who have researched special topics, teacher examination of individual students, and, as described further in Part Two, examination of individual students by a committee of adults. In these situations, students are expected to explain and justify their conclusions, first through an initial statement, and then by elaboration in response to further questions posed by the examiners. As in writing and oracy assessment, specific criteria for evaluation can be articulated and probably scored reliably.

Assessment through extended written and oral discourse raises a number of logistical issues that cannot be solved here. Progress in this direction could

be made, however, by assessing fewer students at a time and staging the assessments throughout a course of study. A teacher who has five classes per day and 125 students could administer a good essay or oral examination to only three students per class each week. In nine weeks, all 125 students could be tested. If this were to occur in each class in each of the students' four main subjects, once a semester or twice a year (e.g., one major essay and one oral exam), it could make a major impact on students' achievement.

2. NAEP Exercises

The following test exercises, drawn from the National Assessment of Educational Progress's Pilot Study of Higher-Order Thinking Skills Assessment Techniques in Science and Mathematics (NAEP, 1987), illustrate that sometimes it is possible to assess depth of understanding without requiring students to produce extensive written or oral statements.

1. Triathlon: Interpreting Data

Students are required by this paper and pencil task to examine data about five children competing in three athletic events and decide which of the five children would be the all-around winner. Students must devise their own approach for computing and interpreting the data and explain why they selected a particular "winner." Students must be careful in their interpretation, because the lower scores are better in the 50-yard dash, while the converse is true in the frisbee toss and weight lift.

Student Assessment Sheet

Joe, Sarah, Jose, Zabi, and Kim decided to hold their own Olympics after watching the Olympics on TV. They needed to decide what events to have at their Olympics. Joe and Jose wanted a weight lift and a frisbee toss event. Sarah, Zabi, and Kim thought running a race would be fun. The children decided to have all three events. They also decided to make each event of the same importance.

They held their Olympics one day after school. The children's parents were the judges and kept the children's scores on each of the events.

The children's scores for each of the events are listed below:

Child's Name	*Frisbee Toss*	*Weight Lift*	*50-Yard Dash*
Joe	40 yards	205 pounds	9.5 seconds
Jose	30 yards	170 pounds	8.0 seconds
Kim	45 yards	130 pounds	9.0 seconds
Sarah	28 yards	120 pounds	7.6 seconds
Zabi	48 yards	140 pounds	8.3 seconds

16

Record Findings

(A) Who would be the all-around winner?

_____

(B) Explain how you decided who would be the all-around winner. Be sure to show all your work.

Account for Findings

_____

_____

_____

_____

Heart Rate and Exercise: Designing an Experiment

Students design an experiment to determine the effects of exercise on heart rate. Students need to identify the variables to be manipulated, specify what needs to be measured, and describe how the measurements should be made to provide reliable results. This exercise can assess students' understanding and planning of scientific investigations when actual experimentation in a classroom or assessment setting is difficult.

Student Assessment Sheet

Usually your heart beats regularly at a normal rate when you are at rest. Suppose someone asks you the following questions:

- Does your heart rate go up or down when you exercise?
- How much does your heart rate change when you exercise?
- How long does the effect last?

Think about what you would do to find answers to the questions above. What type of experiment would you design to answer the questions? Assume that you have the following equipment available to use: an instrument to measure your heart rate (such as a pulse meter), a stop watch, and some graph paper. Briefly describe how you might go about finding answers to these questions.

Describe Experiment

_____

_____

_____

_____

_____

The NAEP tasks include criteria for scoring the responses. For instance, in the Heart Rate and Exercise task, performance is rated on a 0-6 scale. A 6 is awarded for a description that includes all the essential elements for a successful experiment: a baseline measurement (at-rest heart rate); timed exercise; heart rate measured immediately after exercise; and repeated measurements of heart rate over a set period of time until normal. A 1 is awarded for an irrelevant or meaningless description of an experiment or a very incomplete experiment that does not go beyond a mention of exercise. A 0 is given for no

response. One additional point is awarded for any indication of a need to repeat trials in the experiment. Two additional points are awarded for statements indicating the value of repeating the experiment using different durations or intensities of exercise.

The Triathlon is scored in a similar fashion. A 4 is given for an accurate ranking of the children's performance on each event, for citing Zabi as the overall winner, and explaining the results. At the lower end, 1 is given for a selection of an overall winner with irrelevant or non-quantitative calculations or with no explanation; 0 for no response. More specific scoring criteria for each of these assessment tasks are described in NAEP (1987).

3. Alverno College's In-Basket Exercise

The in-basket exercise of the outcomes-based assessment program at Alverno College in Milwaukee poses a variety of problem-solving tasks that require on-the-spot analysis, synthesis, and evaluation (Loacker et al., 1984: 157).

Typically the in-basket exercise requires students to adopt a professional role (manager, secretary, board member, etc.) and describe how they would respond when a decision must be made under realistic time and information constraints. Information is provided in memos, dossiers, and reports, but students must also rely on their prior knowledge about the roles of others involved in the situation.

The students are asked to tell or write about what they would do and also to create or supply whatever written responses (for instance, letters or memos) might be called for. These products are then assessed according to specific criteria. The following examples illustrate a range of in-basket exercises.

1. The student is vice-chair of a school board subcommittee established to deal with complaints about censorship of curriculum materials. The chair of the committee is out of town on an emergency, and the vice-chair must respond to immediate demands, including requests from a coalition of book-banning community activists to establish a formal textbook reviewing committee, a telephone call from a newspaper reporter, and a letter from a state legislator inquiring into the controversy.
2. The student is a teacher with an afternoon devoted to professional activities. The teacher receives a set of communications from the principal instructing him or her to write a diagnostic report based on test and behavioral information about a student, a lesson plan for a substitute teacher, a communication to a parent with a complaint, and a recommendation to the principal regarding a decision to make rules regulating use of a student commons area.
3. A student is given a brief history of and the purposes and programs of an urban cultural center, and, as a newly hired publications specialist, asked to edit an article and reduce its length by a third; handle an irate citizen unhappy with the center's service; write an editorial in response to a newspaper article that quotes the irate citizen and identifies him as an important member of a

Citizens for Tax Reform group campaigning for cuts in government spending; and prepare an outline for a speech to a college class in technical writing.

These exercises are designed to draw on information and experiences from particular courses the students have taken. Students are expected to integrate previous learning with new information provided in the exercise, to adapt their communications to the perspectives and interests of the relevant audiences, and to develop priorities and make decisions with limited information and time.

Performance on the in-basket exercises is usually evaluated by panels of judges. Community professionals with expertise in the roles to be simulated often participate in the assessment process. Where the in-basket exercise calls for oral presentations, audio and video recordings are used to enhance the reliability and precision of assessment. The exercises are also taped to permit students to review and learn from their performance.

The criteria used to assess performance include initiative, adaptability, assertiveness, persuasiveness, problem analysis, decisiveness, and efficiency. Students receive general feedback on their performance, along with statements indicating areas of strength and weakness related to specific criteria. This helps students and their teachers set new learning goals for future work.

4. Frontenac High School's Assessment of Technological Studies

Students in the technological studies program at Frontenac High School in Kingston, Ontario, spend most of their class time designing, constructing, and repairing things, and solving practical problems related to auto mechanics, woodworking, machine shop, drafting, and electrical studies. Until recently, however, these skills were assessed only through paper and pencil tests, a frustrating contradiction for both students and staff. A new procedure has dramatically improved the validity and usefulness of assessment in the technological studies program.

Half the final exam for each course in the technological studies department is a conventional written test. It assesses knowledge of structural properties of materials, principles of design, mechanical and electrical processes, names of tools and equipment, and safety precautions. In addition to the written final exam, students must also demonstrate their ability to use their technical knowledge with a hands-on performance.

For instance, the written test in electrical studies covers, among other topics, Ohm's law and theories of parallel circuitry. In the demonstrations, students are required to create an electrical device based on a wiring diagram. In a recent exam they had to construct an alarm unit to warn a driver that the car's headlights are on after the ignition has been turned off.

Assessment tasks are standardized within each shop, but the procedure varies from shop to shop depending on the contingencies of the task, class size, and the instructional priorities of the shop teacher. For the electrical studies

exam, each student is provided with a work space, a wiring diagram, and a supply of tools and materials. The two-hour time limit that is imposed approximates actual employment conditions in which one is under some pressure to work at a steady pace. How well the alarm unit functions is the main criterion for success, but students are also graded on neatness and precision of work, speed, and safety.

As another example, a used car is brought into the auto mechanics shop. Each student, acting as a consultant to a hypothetical buyer, must identify mechanical problems and make recommendations. The instructor uses a checklist to assess the thoroughness and accuracy of the student's diagnosis and recommendations.

In the technological studies program, this approach is used mainly to assign course-end grades, but it has also provided useful feedback to teachers and students on the process of teaching and learning these skills.

Specific Curriculum Content

The high school curriculum focuses heavily on knowledge of facts, concepts, generalizations, and theories in specific subjects. While the examples above have been organized under the more general competencies of language performance and problem-solving skills, it is important to recognize that success on many of these tasks also requires understanding of specific subject matter frequently taught in high schools.

In the examples of the holistic and primary trait scoring of writing samples and the oracy exercises, the criteria appear to be "content-free," but this is a misconception. Research has shown that writing or speaking clearly requires extensive conceptual knowledge about the meanings of words shared by one's audience.[6]

Assessments that present students with open-ended communication and problem-solving tasks can evaluate student understanding of topical knowledge in two ways. First, the task itself can be designed to focus on problems unique to particular subjects, as was illustrated in the oracy tasks about the garden spider and the human brain, the essay exam about acid rain, the NAEP exercises about empirical inquiry, the Alverno task about censorship (some knowledge of the Constitution would be required), and the Frontenac tasks for electrical studies. Similarly, general writing assessments could require writing about particular subjects in the curriculum.

Second, when appropriate, criteria intended to assess understanding of specific content can be added to the scoring process. Such criteria are included in the essay exam on acid rain and the NAEP exercises. All the examples, however, could require that students incorporate into their responses a set of specific facts, concepts, and theories from the subjects studied.

---

[6]For a summary of some of this research, see Hirsch (1987).

Specific curriculum content can also be assessed through short answer and multiple choice questions. As commonly used, these questions usually fail to assess authentic forms of achievement, but if creatively constructed, they can assess aspects of disciplined inquiry and integration of knowledge.

Discrete Competencies In Perspective

The movement for accountability has expanded competency testing at the school, district, and state levels. Practitioners and researchers alike recognize the double-edged character of competency testing.[7] There is a need to maintain uniform standards and ensure mastery of explicit skills and knowledge, but competency testing can trivialize skills and knowledge, it can produce arbitrary cut-offs between passing and failing, and, if overly centralized, it can suppress creative curriculum and teaching.

Although familiar multiple choice and short answer tests produce quantitative indicators, they cannot assess student production of discourse, things, or performances, and their format works against the assessment of depth of understanding and integration of knowledge. In contrast, the examples in this section show a range of methods that permit quantitative assessment of discrete competencies under standard conditions and that also meet these criteria for authenticity.

Part Two: Exhibitions

Discrete competencies are usually assessed within the confines of schools. In contrast, exhibitions often involve production of discourse, things, and performances for the public. Exhibitions also usually require integration of a broad range of competencies and considerable student initiative and responsibility in carrying out a project. Such projects pose major challenges consistent with the philosophy of the "Walkabout" proposed by Gibbons (1974).

The Walkabout is an Australian aborigine rite of passage to adulthood in which the adolescent must survive alone in the wilderness for several months. Gibbons proposed that this spirit of personal challenge and risk-taking be applied to schooling. Initially he suggested that curriculum and assessment be based on challenges in five main areas: adventure, creativity, service, practical skill, and logical inquiry. Several high schools have adapted these ideas, and Gibbons has developed them further.[8]

---

[7]For more information on student competency testing see Klein (1983-84).

[8]See Gibbons (1976, 1984). The latter reference is accompanied by seven other articles describing the development of Walkabout or "challenge education" in schools since 1974.

Passages at Jefferson County Open High School (JCOHS)

JCOHS is a small public high school in Evergreen, Colo., west of Denver. It is a typical small high school in terms of per pupil expenditures, student-staff ratio, and instruction in some traditional academic subjects, but in most other respects, the school departs markedly from traditional practice. Daily attendance is not required, except for a nine-week orientation that introduces new students to the responsibilities of self-directed learning and helps them select an adviser (teacher). There are no required courses, and neither letter grades nor credits are used. In fact, the majority of student time is probably spent outside classrooms.

The Program

The program is guided by the belief that didactic classroom instruction is only one source of learning and growth; other sources are manual labor, public service, social interaction, individual reflection, and direct experience with diverse environments. Students learn science, history, algebra, and other traditional academic subjects from teachers, community professionals, and sometimes from other students who have expertise to share. They read, write, and do assignments in study rooms equipped with instructional resources. Students do committee work, participate in school governance, and plan schedules and trips. They work in the cafeteria, in the administrative offices, on the school grounds, and at other sites.

And, they travel. Two large, school-owned vans make 10 to 15 trips throughout North America each year. The trips range from several days to two weeks and have included river expeditions, wilderness backpacking, visits to ancient ruins in Mexico, and underwater exploration in the ocean.

Students at JCOHS are involved in shaping and evaluating their own learning. At the beginning of each year, each student, working closely with an adviser, develops an Individual Learning Plan that states what the goals are, how they will be achieved, and what courses or activities will be pursued. A parent conference ensures that clear expectations are shared by the student, the staff members, and parents. In bi-weekly meetings with the adviser, the student discusses and shows evidence of progress, and determines if plans should be modified.

The student receives formal recognition for completing the Individual Learning Plan after the adviser receives written evaluations from the student about each completed activity with a corresponding response by the teacher or other person in charge of the activity.

After completing this phase, which can take from one semester to several years, the student begins the "passages" which, as described below, are the culminating challenges to demonstrate the diverse competencies required for graduation. The final requirement for graduation is a well-organized, 20 to

30-page written "transcript" that summarizes and interprets the learning that has been documented through evaluations and passages.

Flexibility in the organization of courses, instruction, and related activities is critical to the program, and is made easier by the school's small enrollment—about 250 students. Each certified staff member is responsible for 12 to 20 students and spends 30 percent to 50 percent of his or her time advising, which fosters in the adviser a close personal knowledge of the abilities, interests, and progress of the students. Weekly schoolwide governance meetings provide an open forum for everyone to discuss issues and make decisions. Courses, community experiences, and other learning activities are initially structured in blocks of several weeks to several months. But course offerings and school-sponsored activities can be changed through the governance meetings, and Individual Learning Plans can be changed in consultation with the advisers.

Passages at JCOHS

The "passages" are designed to demonstrate competence in six broad areas: practical skills, creativity, adventure, career exploration, logical inquiry, and global awareness/volunteer service.

Students learn about the passages by meeting in advisory groups, participating on passage committees, and observing others fulfilling their passage requirements. However, students cannot begin their own personal passages until they demonstrate the requisite level of independence and academic competence to their advisers.

A written proposal is prepared according to several guidelines. Students:

1. State their proposal as a challenge, being specific and stating the kind of performance and the level of performance they will pursue.
2. Outline the preparation they will need—training, practice, information gathering, and so on.
3. List the resources they will need—equipment, people, work space, transportation, materials, money, etc.
4. Indicate what is the greatest obstacle they expect to encounter.
5. Explain what positive sources they can draw on to overcome this obstacle.
6. Identify their first step in launching the passage.
7. List the other steps that lead to the completion of the passage.
8. Identify what form their presentation or demonstration will take upon completion of the passage.
9. Indicate their proposed date of completion.

The student discusses the written proposal with the adviser, changes it if necessary, and then presents it to a staff member chosen on the basis of expertise as a passage consultant. The proposal may require further modification. It is then presented to the student's passages committee made up of the adviser, parent(s), principal, a student experienced with passages, another

student, and the passage consultant. When the committee approves, the student begins the passage.

Passages can last from two months to two years. Some examples illustrate the many different activities on which passages have been built:

- Running the lunch room
- Doing construction and maintenance work around school
- Helping an elementary school physical education teacher supervise the playground and take care of younger children
- Doing library research on AIDS and working on the Denver AIDS hotline
- Working with a Denver energy cooperative to weatherize homes and assist the poor to surmount home energy problems
- Teaching a class at JCOHS
- Working as an apprentice to a chemist involved in research on the effects of carbon dioxide on the atmospheric ozone layer
- Volunteering at the local chapter of the Audubon Society and participating in research on wetlands ecology.

When the passage is complete, the student writes an extended evaluation of the experience and makes a presentation to his or her passage committee. The purpose of the presentation is to document the effort expended, the resulting achievements, and the ways the student has benefited intellectually and emotionally. Slides or pictures can be displayed, along with products, testimonial letters, and other evidence. Successful completion of all six passages must be demonstrated for graduation. The committee reviews the written evaluation and presentation with the student, and if the documentation is considered acceptable, they celebrate the completion of the passage.

A final requirement—one students have come to call "the seventh passage"—is a written summative evaluation of the high school experience. This is a reflective discussion of what they have done and learned. The document is used in place of the traditional high school transcript.

Walden III's Rite of Passage Experience (ROPE)

This program uses exhibitions in a major senior year project to guide students through a process of "pulling it all together" and to evaluate their achievement on graduation requirements.

Walden III High School, Racine, Wis., has a graduation requirement called the "Rite of Passage Experience," (ROPE). All seniors must demonstrate mastery in 15 areas of knowledge and competence by completing a portfolio, project, and 15 presentations before a ROPE committee consisting of staff members (including the student's homeroom teacher), a student, and an adult from the community. Nine of the presentations are based on the materials in the portfolio and the project; the remaining 6 presentations are developed specially for the presentation process itself.

Procedure

The Portfolio. The portfolio, developed during the first semester of the senior year, is intended to be "a reflection and analysis of the graduating senior's own life and times."[9] Its requirements are:

1. *A written autobiography*, descriptive, introspective, and analytical. School records and other indicators of participation may be included.

2. *A reflection on work*, including an analysis of the significance of the work experiences for the graduating senior's life. A resume can be included.

3. *Two letters of recommendation* (at minimum) from any sources chosen by the student.

4. *A reading record* including a bibliography, annotated if desired, and two mini-book reports. Reading test scores may be included.

5. *An essay on ethics* exhibiting contemplation of the subject and describing the student's own ethical code.

6. *An artistic product or written report* on art and an *essay on artistic standards* for judging quality in a chosen area of art.

7. *A written report analyzing mass media*: who or what controls mass media, toward what ends, and with what effects. Evidence of experience with mass media may be included.

8. *A written summary and evaluation of the student's coursework in science/technology; a written description of a scientific experiment* illustrating the application of the scientific method; *an analytical essay* (with examples) on social consequences of science and technology; *and, an essay on the nature and use of computers* in modern society.

The Project. Every graduating senior must write a library research-based paper that analyzes an event, set of events, or theme in American history. A national comparative approach can be used in the analysis. The student must be prepared to field questions about both the paper and an overview of American history during the presentations, which are given in the second semester of the senior year.

The Presentations. Each of the above eight components of the portfolio, plus the project, must be presented orally and in writing to the ROPE committee.

Six additional oral presentations are also required. However, there are no written reports or new products required by the committee. Supporting documents or other forms of evidence may be used. Assessment of proficiency is based on the demonstration of knowledge and skills during the presentations in each of the following areas:

9. *Mathematics knowledge and skills* should be demonstrated by a combination of course evaluations, test results, and worksheets presented before the committee; and by the ability to competently field mathematics questions asked during the demonstration.

---

[9]The following draws from the 1984 student handbook, "Walden III's Rite of Passage Experience," by Tom Feeney, a teacher at the school.

10. *Knowledge of American government* should be demonstrated by discussion of the purpose of government; the individual's relationship to the state; the ideals, functions, and problems of American political institutions; and selected contemporary issues and political events. Supporting materials can be used.

11. *The personal proficiency* demonstration requires the student to think about and organize a presentation about the requirements of adult living in our society in terms of personal fulfillment, social skills, and practical competencies; and to discuss his or her own strengths and weaknesses in everyday living skills (health, home economics, mechanics, etc.) and interpersonal relations.

12. *Knowledge of geography* should be demonstrated in a presentation that covers the basic principles and questions of the discipline; identification of basic landforms, places, and names; and the scientific and social significance of geographical information.

13. Evidence of the graduating senior's successful *completion of a physical challenge* must be presented to the ROPE committee.

14. *A demonstration of competency in English* (written and spoken) is provided in virtually all the portfolio and project requirements. These, and any additional evidence the graduating senior may wish to present to the committee, fulfill the requirements of the presentation in the English competency area.

Timeline

At the beginning of their senior year, all students enroll in a semester-length ROPE class. In consultation with the ROPE supervisor, they select the teachers, outside adult, and student to be their ROPE committee members. In the ROPE class, students receive instruction from regular subject-matter teachers, and supervision and guidance in the development of their portfolio and project. They also complete a portion of the work for these tasks. The ROPE instructor monitors, advises, reprimands, and keeps students on schedule. Both the portfolio and the project must be completed and handed in to the instructor by the end of first semester.

During the second semester, all seniors make oral presentations to their committees about the tasks of their portfolio, their project in American history, and on the other areas in which demonstrations of proficiency are required. Presentations usually last an hour to an hour and a half. The number of presentations students give varies, but most complete the requirements with five to nine separate appearances before their committee. (More than one requirement may be completed in a single appearance.)

Evaluation

The ROPE instructor monitors and ensures overall presentability of the portfolio and project, but does not grade the students. The ROPE committees evaluate the portfolios and projects on a pass-fail basis and each of the oral

presentations on an A-F grading scale. Each committee determines its own standards for grading, but generally, quality and students' serious application of themselves are important. In addition, the work must be well-organized, grammatically correct, neat, and reflect the English proficiency level of a high school graduate as judged by the student's committee.

To earn a diploma, a student must receive a passing grade in at least 12 of the required areas of the ROPE, which also meet district requirements for proficiency in math, government, reading, and English.

Community-Based Learning and Learning Contracts at "Learning Unlimited"

This program uses learning contracts to guide and hold students accountable for community-based learning. The contracts provide criteria applied both to exhibitions and to more conventional paper and pencil tests.

At North Central High School in Indianapolis, Ind., Learning Unlimited is a school-within-a-school that enrolls about 400 students and offers 22 semester-length academic courses. Students take traditional academic courses, but spend much time learning on their own and participating in the community. Through contractual agreements between student, teacher, and community resource persons, Learning Unlimited strikes a balance between a uniform academic curriculum and an individualized student-directed learning program.

The Curriculum: Courses and Community Work

Each Learning Unlimited course is organized around a set of general curriculum goals and more specific objectives developed by the teachers in the high school. For example, General Goal 11 for history states that the student must be able to: "(1) cite two reasons for the development of the Cold War; (2) define collective security and identify three areas in the world where America is committed to this concept; (3) compare and contrast our involvement in Korea and Vietnam . . ."

Each student is given a list of the course objectives for the semester and has the option of attending class regularly and following recommended readings and assignments, or creating an individualized set of tasks and objectives aimed at the overall course goals and objectives. In either case, a formal contract is signed by student and teacher. At the end of every course a required final exam covers the course objectives.

In addition to regular course requirements, all Learning Unlimited students must complete a minimum of 24 hours of community-based learning experiences each semester in both their junior and senior years. Most students

also participate in a freshman/sophomore field trip, and all upper level students are encouraged—but not required—to participate in "intensives"—several day to several week trips to other cities or regions.

The Learning Contract

Each student's learning is guided by and assessed according to the individual learning contract. A contract covers six weeks and spells out goals, competencies, conference dates with the teacher, activities, a timetable, and criteria for evaluation. It is signed by the student, teacher, and parent. Each contract is developed through consultation and negotiation with the course instructor.

To help students specify academic objectives for their contracts and to provide a common curriculum, Learning Unlimited teachers formulated 12 general academic goals and a more extensive list of specific academic competencies. In developing a contract, a student is free to draw on these competencies. A list of readings and supplementary materials connected to the academic competencies is also available for students to use in the learning contracts.

The community-based learning experiences are also incorporated in the contracts. More than 100 community resource sites participate regularly in the program, and agreements between the student and the community resource person are established in separate contracts. One student volunteered four to five hours a week working with the professional organizers of the Pan Am games in Indianapolis. Other students have worked at the zoo, at businesses, at fire and police stations, and as aides at day care centers, hospitals, and nursing homes.

Students must indicate on their learning contract what they will do and what they expect to achieve. Most often the community experiences result in some form of written product, such as a journalistic account of a community issue, a sociological or political analysis based on data from participation in a community agency or service, or a personal reflection on an experience presented in a diary-like format. Slides, physical products, and anything else that demonstrates thoughtfulness and achievement resulting from the community experience are encouraged.

A single community project can serve as a resource for more than one learning contract. For instance, from the Pan Am games project, the student developed a slide show presentation for his photography class, a paper on the international issues involved in the Pan Am games for his government class, and a journal of his daily experiences and observations for his English class.

Evaluation

Each six weeks, grades are determined during a conference between the student and the teacher. Final course grades are determined by final exams and students' performance in relation to their learning contracts.

Students are also evaluated by the community resource person who signs the student's time card and fills out a standard evaluation form. Criteria are dependability, attitude, effort, and extent of fulfillment of the community project contract. The form also asks the resource person to assign a letter grade on overall performance and to assess the quality of the overall experience for the student.

This same general contractual process occurs for the shorter field trips and the longer, more in-depth "intensives." That is, students must reflect on their experiences and communicate their thoughts to the class or to some other audience in accordance with the timeline and quality specifications of the learning contract.

Exhibitions in Perspective

The programs in which exhibitions occupy a central role emphasize the student as an independent worker and the teacher as a resource or coach. Teachers and principals in these schools note that students often have difficulty adjusting to the independence. Their struggle can be eased through special preparation courses and regular teacher-student conferences. Teachers may also need special help in becoming less a dispenser of knowledge and more a coach, adviser, and counselor with a better knowledge of the student's capabilities and interests.

Exhibitions need not be confined to small high schools or to schools dedicated to nontraditional forms of teaching, learning, and assessment. Within conventional high school courses, teachers can (and do) assign major projects, either as special unit activities or as end-of-course demonstrations, that could qualify as exhibitions (e.g., production of videotapes, public debates, publication of oral histories). Exhibitions within courses can bring special life to instruction.

Since exhibitions do not rely exclusively on traditional tests and offer significant recognition for mastery in arts, crafts, academic and innovative endeavors, they have unique potential for engaging otherwise alienated students.

Public exhibitions of mastery can clarify standards of achievement and celebrate the ideal of competence for a broad range of students. More than any other form of assessment, exhibitions of mastery are most likely to fulfill all the criteria for authentic achievement: disciplined inquiry, integration of knowledge, and value beyond evaluation.

This form of assessment requires extra staff work and some changes in class scheduling and course structure. It is time-consuming to carefully manage and judge exhibitions. However, on the basis of reports from faculty members and students, the high level of student participation in learning and the clarity of educational purposes fostered by these approaches makes them well worth considering.

Part Three: Portfolios And Profiles

Portfolios and profiles furnish a broad, often longitudinal, portrait of individual performance in several dimensions. Each can offer multiple authentic indicators of achievement. Portfolios are intended to give comprehensive, cumulative portraits, but not necessarily on standardized indicators. Profiles emphasize teachers' ratings on scales of diverse competencies or student characteristics such as perseverance.

Portfolios

A portfolio is a file or folder containing a variety of information that documents a student's experiences and accomplishments. The portfolio can contain summary descriptions of accomplishments, official records, and diary items.

Summarized descriptions of accomplishments can include samples of writing; audio, video, and photographic recordings of performances and projects; and testimonies from authorities about the quality of student work. Experiential learning that is the grist for exhibitions can also be summarized. Students have described such examples as designing a mechanical device to help a handicapped friend; organizing a talent show; taking the responsibility of arranging meetings with visitors to the school and showing them around; forming a group to grow a garden and selling the produce to earn income; and participating in charity fundraisers.[10]

A variety of formal records are usually included also: a curriculum transcript, scores on standardized tests or other examinations taken during high school, evidence of membership and participation in school clubs or academic events, a list of awards or any other distinctions, and letters of recommendation.

To encourage students to reflect on their learning, they may be required to keep a diary. The portfolio will be enhanced by including excerpts from the diary that illustrate the student's view of his or her intellectual and emotional development.

A portfolio gains legitimacy if its contents are validated by appropriate authorities, such as teachers, guidance counselors, or community representatives. The learning contracts of Learning Unlimited require signatures of an adult representative responsible for overseeing the community learning experience, the student's parent or guardian, and the teacher supervising the experi-

[10]See Burgess & Adams (1985).

ence. The Walden III portfolio, part of the ROPE program to certify graduation, is validated by experts in the disciplines represented in the portfolio, someone with more general knowledge of the candidate's character and abilities, and an official representative of the school.[11]

Profiles

Unlike portfolios, profiles are not created by individual students and do not contain actual samples of work. Rather, profiles are forms that teachers, students, and sometimes parents fill out with ratings and summary judgments or descriptions of achievement.

In England, a profile system has been developed with several goals in mind: to convey a rich variety of information about the interests, character, accomplishments, and academic proficiency of individual students; to allow comparative judgments across time and among peers; and to involve minimal record keeping.[12]

Teachers at the Wootton Bassett School, Swindon, England, developed an assessment system that is both individualized and standardized, and that provides a more comprehensive picture of student achievement than do grades or test scores. Although this approach was designed for science, it could be adapted for other subjects.

At the end of every unit (before the unit exam is graded), each student's performance in effort, presentation, communication, and research is recorded on an assessment card.

Student performance is scored on a 1-4 scale for each of the four major criteria. Each criterion has specific standards for these ratings. For example:

RESEARCH:
1—Regularly shows originality of thought or action beyond that taught or set.
2—Now and again shows qualities as in 1.
3—Rarely shows qualities as in 1.
4—For their ability they are doing as required, yet have shown no originality or initiative in the topic. (This person could still be top of the class in other respects.)

Cards for each student go into individual student files, thus providing a detailed and comprehensive record of achievement in individual subjects.

A national commission of educators and private citizens in Scotland was charged with overseeing the development of an approach to assessment that

[11]For more information on the use of portfolios, see Committee for the Assessment of Experiential Learning (1975) and Forrest (1975). Interesting work on portfolios in the teaching and assessment of art is taking place in the The Arts PROPEL project, a collaborative effort between Educational Testing Service, Harvard Project Zero, and the Pittsburgh Public Schools funded by the Rockefeller Foundation. See PROPEL (1987).

[12]The examples are drawn from Burgess and Adams (1980).

"would be equally applicable to all pupils; which would gather teachers' knowledge of pupils' many different skills, characteristics and achievements across the whole range of the curriculum, both formal and informal; which would, with a minimum of clerical demands, provide a basis for continuing in-school guidance, culminating in a relevant and useful school-leaving report for all pupils." The following approach was developed by the Scottish Council for Research in Education working with teachers in a variety of comprehensive high schools over a three-year period.

The assessment system has three parts:

- A *Class Assessment Sheet* to be filled in by the teacher at the end of the course (or more frequently, if appropriate). The teacher rates each student on a one to four criterion-referenced scale in the relevant skills and performance categories. The content of the individual performance categories are to be described by the teacher according to course objectives and assessment needs.
- *Pupil Profile* cards for each student which contain the skills and performance categories including the teacher's ratings for each student in the class. Each card has an additional space for comments.
- Each student's record of achievement summarized on a four-page *School Leaving Report*. The cover page provides biographical information. Page two, "Skills," reproduces the eight individual categories composing the larger skills criterion used on the *Class Assessment Sheet* (Fig. 3.3). A summative rating is given for each of the categories, along with a description of the criterion-referents. Page three, "Subject/Activity Assessment," gives detailed course information and reproduces the composite grade, perseverance, and enterprise categories included in the performance criterion on the *Class Assessment Sheet*. The back page provides space for recognition of individual achievements and comments on other personal strengths. It is the responsibility of some adult(s) who knows the student well (homeroom teacher, guidance counselor) to develop the leaving report from the information contained in the student's *Profile*.

These three parts of the pupil profile system can contribute to several assessment goals. Teachers keep the *Class Assessment Sheet* as a record of achievement for their classes. The individual *Pupil Profile* cards create a detailed individualized longitudinal record of pupil achievement and progress. These cards can be used for conventional grading purposes as well as for diagnostic and guidance purposes. The *School Leaving Report* provides an informative summary of the pupil's high school achievements. Relatively detailed information that is both comparative and idiosyncratic is provided on course background, skills, achievements, and character.

The major advantage of portfolios and profiles is their recognition of multiple indicators of individual achievement. They do require more complex and comprehensive record keeping than do grades and test scores, and there is some risk that well-intentioned efforts to develop multiple indicators will lead to a proliferation of arbitrary and unnecessary criteria for achievement. If we

are interested in more comprehensive indicators of student achievement, however, portfolios and profiles offer promising opportunities.

Summary

What kinds of information might be gathered about a student that conveys both to the student and to others a valid indication of the student's mastery of authentic academic challenges? The diverse approaches described in this chapter fall into three main categories, and each can make an important contribution to a comprehensive scheme.

Testing of more authentic competencies in large scale assessments may appear to be too costly, compared to the costs of conventional standardized multiple choice tests. But there is reason to question this assumption if one considers the total costs of each approach that involve test development, test administration, scoring, and reporting. Standardized tests entail a tremendous investment in development of individual test items, with relatively lower costs in scoring. In contrast, more authentic approaches involve substantial scoring costs and lower costs for development.[13]

Testing discrete competencies informs students about how well they meet public standards for the performance of important, specific skills in language use, problem solving, and mastery of specific subjects.

The main strength of the exhibition is the opportunity it provides students to demonstrate knowledge in ways that have meaning to others. Such demonstrations are often exciting, not only because they have public value beyond testing, but because they allow students to integrate knowledge in unique ways.

Finally, portfolios and profiles present comprehensive summaries of a variety of student accomplishments, thus giving both the student and the public a more global, elaborate record of achievement.

We now turn to the problem of assessing an entire school. The task is complex, because many of the indicators we have discussed cannot easily be aggregated to the school level, and even if they could, meaningful school assessment is more complex than simply averaging indicators of individual student performance.

---

[13]Building on the work of the Assessment Performance Unit (Department of Education and Science, Elizabeth House, York Rd., London, SE1 7PH), a recent project in Connecticut completed a hands-on assessment in science of 900 students in grades 4, 8, 11. Students in 8th and 11th grade had to design, conduct, and record the results of an experiment and also demonstrate competence with scientific equipment (e.g. operating a microscope, triple beam balance, and wiring an electric circuit). Individual students' work was assessed on-the-spot by adult monitors who assigned multiple scores and also made a holistic judgment of work quality. The total cost of this project for 900 students was about $6 per student. Extensive work in writing assessment also suggests that more authentic assessments can be feasible from a cost standpoint (Newmann, 1988).

Figure 3.3

SKILLS

LISTENING

- Acts independently and intelligently on complex verbal instructions ☐
- Can interpret and act on most complex instructions ☒
- Can interpret and act on straightforward instructions ☐
- Can carry out simple instructions with supervision ☐

SPEAKING

- Can debate a point of view ☐
- Can make a clear and accurate oral report ☐
- Can describe events orally ☐
- Can communicate adequately at conversation level ☒

READING

- Understands all appropriate written material ☐
- Understands the content and implications of most writing if simply expressed ☒
- Understands uncomplicated ideas expressed in simple language ☐
- Can read most everyday information such as notices or simple instructions ☐

WRITING

- Can argue a point of view in writing ☐
- Can write a clear and accurate report ☐
- Can write a simple account or letter ☒
- Can write simple messages and instructions ☐

VISUAL UNDERSTANDING AND EXPRESSION

- Can communicate complex visual concepts readily and appropriately ☐
- Can give a clear explanation by sketches and diagrams ☐
- Can interpret a variety of visual displays such as graphs or train timetables ☒
- Can interpret single visual displays such as roadsigns or outline maps ☐

USE OF NUMBER

- Quick and accurate in complicated or unfamiliar calculations ☐
- Can do familiar or straightforward calculations, more slowly if complex ☒
- Can handle routine calculations with practice ☐
- Can do simple whole number calculations such as giving change ☐

PHYSICAL COORDINATION

- A natural flair for complex tasks ☐
- Mastery of a wide variety of movements ☐
- Can perform satisfactorily most everyday movements ☒
- Can perform single physical skills such as lifting or climbing ☐

MANUAL DEXTERITY

- Has fine control of complex tools and equipment ☐
- Satisfactory use of most tools and equipment ☒
- Can achieve simple tasks such as wiring a plug ☐
- Can use simple tools, instruments and machines such as a screwdriver or typewriter ☐

SUBJECT ACTIVITY ASSESSMENT

Curriculum Area	Subjects Studied (includes final year level where relevant)	Years of Study	Achievement	Enterprise (includes flair, creativity)	Perseverance (includes reliability, carefulness)
Aesthetic Subjects	Drawing Music	1-4 1-4	2 2	2 3	1 3
Business Studies					
Community/ Leisure Activities	Social Education	1-4	3	2	3
Crafts	Pottery	3-4	2	1	3
English	English	1-4	2	1	3
Mathematics	Arithmetic	1-4	1	1	2
Other Languages	German	2-4	2	2	3
Outdoor Studies	Outdoor Pursuits	3-4	2	2	3
Physical Education	General	1-4	3	1	3
Science	Biology	3-4	1	2	2
Social Subjects	History	1-4	2	1	3

CHAPTER 3

Assessing Organizational Academic Quality

ASSESSING THE QUALITY or productivity of the school as a whole raises at least two questions: What standards or reference points should be used to judge school success? How can the accomplishments of diverse students be reflected in indicators that convey useful information about the school? This chapter begins with a discussion of guidelines for developing organizational standards. It then presents several examples of indicators of authentic achievement that can be aggregated and interpreted in terms of school standards.

Guidelines for Organizational Standards

Judgments about school performance have little meaning unless performance is described in relation to a standard point of reference. A standard is a set of baseline criteria that have reasonably uniform or common meanings across time and place. For example, claims about the percent of students who demonstrate competence in writing should be based on tasks and criteria for evaluation that are constant for all students who take the test. Claims about reduction in the dropout rate should be based on a uniform procedure for computing the rate from year to year (Williams, 1987).

A set of standards is most meaningful if it provides for longitudinal comparison, comparison between schools, and disaggregation of data within the school, and if it includes indicators responsive to unique school goals. Each of these guidelines is discussed below.

1. *Longitudinal comparisons.* Unless a school has information about student performance at two or more different times, there is no basis for estimating the effect of the school on student performance. Of course, pre- and post-assessments offer no guarantee that observed changes can be attributed to the school program alone. The influence of other factors (e.g., students' personal background, a changing student body, unique community events) may be difficult to distinguish from school effects, but longitudinal data is, nevertheless, necessary. How frequently data should be collected depends on the schools' goals and on the desirability of comparison with other schools during specific time periods. At a minimum, however, it would seem useful to assess a sample of freshmen and seniors each year.

   If pre- and post-standardized test scores are used, it should be recognized that these tests are not designed to measure cognitive growth over time (Heyns, 1978). Their purpose is to assess performance at a particular time, and the non-uniform intervals of standardized test scales make it difficult to know how much knowledge is gained or lost. Essay tests, oral exams, and some of the other examples described in Chapter 2, administered on a pre- and post-basis, may provide more informative, more authentic indicators of growth.

2. *Comparisons between schools.* There are some domains of achievement, especially in reading, writing, speaking, mathematics, and citizenship, that all schools should have an interest in promoting and which the public has a right to expect of all high schools. Schools identified as less successful will be able to target their efforts on specific areas. Schools identified as most successful can serve as sources of inspiration and assistance to the less successful. Of course, care should be taken to compare only schools that have similar curricula and students.

3. *Disaggregated data within the school.* When test scores reduce a school's academic quality to a single number, such as an average score or percentile ranking, they conceal potentially important patterns of variation. Separate student groups within the school, such as college-bound, low-income, or handicapped students may perform quite differently on the test. A single score may not detect important differences in performance on mechanics and organization. Two schools may have the same score yet very different patterns of achievement. Or a program may cause a decrease in the range of scores over a period of years (reducing the number of very high and very low scores) without changing the school's mean score.

The standards for judging performance should, therefore, anticipate the use of disaggregated data and include standard deviations, the use of medians and percentages of students falling within different sections of the distribution, and more specific comparisons between groups of students or domains of content.[1]

4. *Indicators of unique school goals.* To capture unique objectives of the school, the range of indicators can be expanded in four directions.

   Special attention can be given to curriculum-aligned exercises—both those designed for the particular school or district and those produced for national or international use (e.g., subject matter tests in the Advanced Placement or International Baccalaureate programs). Compared to standardized tests of general achievement and ability, curriculum-aligned forms of assessment allow more valid inferences about curriculum and instructional quality.[2] Curriculum-aligned assessments can be developed to meet departmental objectives. For example, the science department may want a special assessment on laboratory skills and the social studies department may wish to assess student ability to explain current issues in terms of historically relevant events.

   In addition to assessing performance, schools may wish to establish standards for student participation in academic coursework and academically demanding cocurricular activities.

   The range of indicators can also be extended by establishing distinct standards and setting special goals for specific groups of students, such as at-risk students or those in a particular curriculum track.

Finally, a range of different quantitative indicators should be considered in selecting standards unique to the school. These can include fixed performance levels (e.g., 70 percent of the students score three or higher in essay writing); longitudinal change rates (e.g., 2 percent reduction in dropout rate); between-school comparative criteria (performing within the top 50 percent of comparable schools); and dispersion of success across the student enrollment (e.g., reduction of variance in mathematics achievement, along with an increase in the mean).

These guidelines should not lead to a vast increase in the amount of student time devoted to test taking. In building indicators of school success, schools can minimize the burden to students and teachers by obtaining data from samples of students rather than from the whole student body.

---

[1]Big city school systems are collecting and reporting increasingly detailed statistical information on test performance. Scores broken down by ethnic and income categories, by schools, grades, and skill areas can give more precise information, but also can be the source of controversy. See the Association for Supervision and Curriculum Development's *Update*, March 1987.

[2]Subject-specific standardized tests avoid some problems, but limitations due to multiple-choice format remain. Recent reviews have raised critical questions about the items on these tests. See for example, Murnane and Raizen's (1988) review of tests in science and mathematics.

Examples of Innovations in Organizational Assessment

Participation Indicators[3]

Significant learning requires effort by the student, but many students can graduate with minimal exertion (Powell, Farrar, and Cohen, 1985). In a sense, the best teachers and schools may be those who inspire students to work harder and to participate actively in school life. Schools that generate high levels of student participation should see this as an indicator of their academic quality. Several types of participation indicators can be used to develop a school profile on academic participation.

1. Rates of attendance, dropout, and disciplinary action.
2. Indicators of academic engagement, such as percentage of seniors taking college placement tests and percentage of seniors enrolled in nationally recognized advanced credit programs.
3. Indicators of participation in cocurricular activities, such as volunteer service in social agencies (hospitals, schools, charitable organizations), support for political parties and campaigns (including registering to vote and voting), helping advocacy groups (consumer protection, civil rights, women's issues), participation in self-help and support groups (drug abuse, teen pregnancy, and parenting), and speaking out as an independent citizen (letters to the editor).
4. Postsecondary achievements, such as percentage of recent seniors enrolled in postsecondary institutions, average freshman college grades of recent graduates, percent of recent graduates employed, number of return visits to the school, or unique accomplishments of graduates.

Some participation indicators, especially awards, may give a meaningful account of the quality of academic achievement, but others, such as attendance rates, convey little about the quality of specific accomplishments. Thus, participation indicators are necessary criteria of academic quality, though insufficient alone. They complement those indicators tied more directly to the content of actual achievements.

---

[3]The following description of school-level indicators and criteria of academic quality draws from school recognition programs sponsored by the California State Department of Education, the United States Office of Education, the Ford Foundation, and the University of Illinois at Chicago. For an extensive list of national contests, activities, and awards programs, see the *NASSP National Advisory List of Contests and Activities*, available from NASSP.

Disaggregating and Reporting

Important information can be added by disaggregating test results to allow for comparisons over time, between schools, between groups of students, and between domains of competence. Innovative reporting of quantitative information, however, cannot compensate for deficiencies in the tests or other indicators. Test items should meet criteria for authenticity such as critical substantive or procedural knowledge in a field, in-depth understanding or integration of knowledge, and, if possible, production of discourse, things, and performances that have value beyond evaluation.

If the intent is to evaluate the effect of school instruction, the items should also be aligned with the actual curriculum. No test can cover everything that has been taught, but the content and skills measured by the test should match those that have been taught by teachers.

1. A Curriculum-Referenced Reporting System

Adlai Stevenson High School (1607 W. Hwy 22, Prairie View, Ill. 60069) bases assessment on results from:

- Its curriculum-referenced testing (CRT) program that covers the great majority of the school's courses and includes some NAEP items for national reference points;
- A primary trait writing assessment of all students at the end of the required sophomore/junior composition course;
- The nationally standardized Scott Foresman and College Board Advanced Placement tests covering all the core curriculum subjects;
- Sophomore (California Achievement Test) and senior level (ACT) tests of general achievement.

The testing program provides information about achievement at multiple levels (student, classroom, school); across all subjects; and in relation to nationally standardized norms, fixed performance criteria developed by the school (i.e., specific statements of learning objectives), and past performance (by using pre-tests/and post-tests in several courses and by comparing present scores to scores from previous years).

The testing program relies heavily on computers, and the school employs a part-time computer specialist to provide technical assistance, coordination, and test analysis. Departments also use their own microcomputers for keeping records of student data and for analyzing test results.

Because tests, syllabi, and lists of learning objectives are stored in computerized files, they can be easily adapted to changing needs and goals via a word processor. Machine-gradable answer sheets are used for testing, and in most cases test results can be returned to teachers in an hour.

The computer software produces a variety of information about student

test performance on easy-to-read tables and charts. Teachers can see which test items were easy or hard for the class and whether there were patterns of mistakes on any questions. At Stevenson, one analysis revealed that students were more likely to answer incorrectly, or choose "none of the above," on questions with relatively long stems and distractors. Some of the longer questions with longer answers were confusing and the construction of some of the items needed to be reviewed. This also stimulated discussion about how to improve students' abilities to handle more complex information processing.

Item analysis can reveal student performance on subtests—groups of questions covering a single topic. For instance, an American history test at Stevenson covers seven topic areas: chronology, government, ideology, foreign policy, political history, economic history, and social history.

While the main purpose of the CRT is to evaluate student performance in relation to teacher-specified learning goals, the CRT test is anchored in national reference points by matching items common to the CRT and a nationally standardized test of American history used at Stevenson.

At Stevenson, the curriculum tests, the nationally normed tests, and the computer data base constitute an integrated testing program. Breakdowns by topics and student categories linked to different norms can help teachers assess strengths and weaknesses in their instruction, identify special patterns of outcomes for particular students, and assign grades. At the end of the year, "accountability reports" to the administration and the community summarize school performance and generate discussions about program improvement.

2. Reporting Writing Achievement

Detailed data about writing achievement can be reported, as shown in the reporting of the writing assessment program of Adams County School District #12, Northglenn, Colo.

Figure 3.1 gives both the criteria for performance and numerical results, using weighted scores to differentiate the importance of different aspects of writing. The maximum score attainable is 100 ("5" on each of the criteria, multiplied by their respective numerical weightings). The pie chart shows the percent of papers with scores from 75 to 100 (2 percent), 50 to 74 (13 percent), 25 to 49 (50 percent), and 0 to 24 (35 percent). Although the mean was 61, note that more than 85 percent of the students scored less than this.

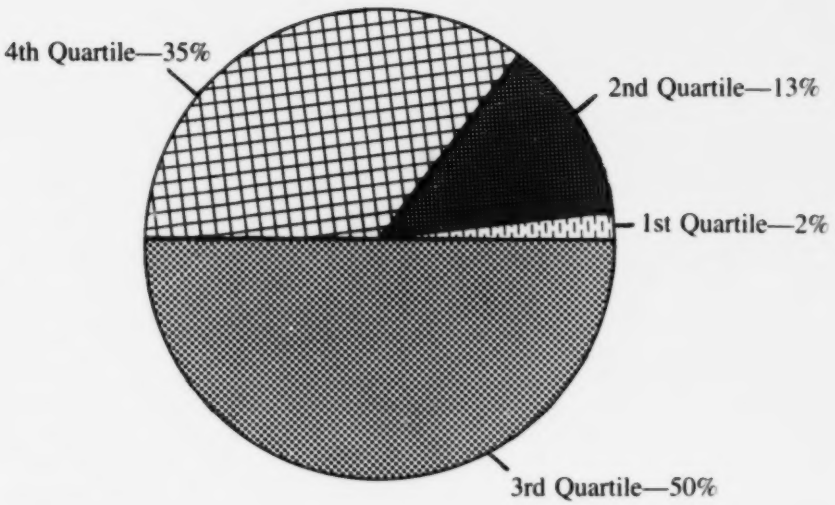
The weighted means and the total score are useful as comparative reference points. For example, scores from individual eleventh grade students or from classes can be compared to the districtwide mean scores for eleventh graders. The districtwide means can also be compared to mean scores from previous years to develop longitudinal data on writing achievement.

Figure 3.1
Eleventh Grade Writing Sample Results

	1	2	3	4	5	
Organization	Little or nothing is written. The essay is disorganized, incoherent, and poorly developed. The essay does not stay on the topic.		The essay is not complete. It lacks an introduction, well-developed body, or conclusion. The coherence and sequence are attempted, but not adequate.		The essay is well-organized. It contains an introductory supporting and concluding paragraph. The essay is coherent, ordered logically, and fully developed.	x6
Sent. Str.	The student writes frequent run-ons or fragments.		The student makes occasional errors in sentence structure. Little variety in sentence length or structure exists.		The sentences are complete and varied in length and structure.	x5
Usage	The student makes frequent errors in word choice and agreement.		The student makes occasional errors in word choice or agreement.		The usage is correct. Word choice is appropriate.	x4
Mechanics	The student makes frequent errors in spelling, punctuation, and capitalization.		The student makes an occasional error in mechanics.		The spelling, capitalization, and punctuation are correct.	x4
Format	The format is sloppy. There are no margins or indentations. Handwriting is inconsistent.		The handwriting, margins, and indentations have occasional inconsistencies—no title or inappropriate title.		The format is correct. The title is appropriate. The handwriting, margins, and indentations are consistent.	x1

	Mean	*Weighted Mean*
Organization	2.91	17.43
Sentence Structure	3.38	16.89
Usage	2.93	11.72
Mechanics	2.90	11.58
Format	3.42	3.42
Total Score		61.04

The eleventh grade total mean is 61.04. The areas of strength are format and sentence structure. The remaining areas, organization, usage, and mechanics, have similar scores. Organization, which carries the greatest considered weight, should be considered the weakest trait.



Source: Adams County School District #12, 11285 Highline Dr., Northglenn, Colo. 80203

3. School and District Statistical Profiles

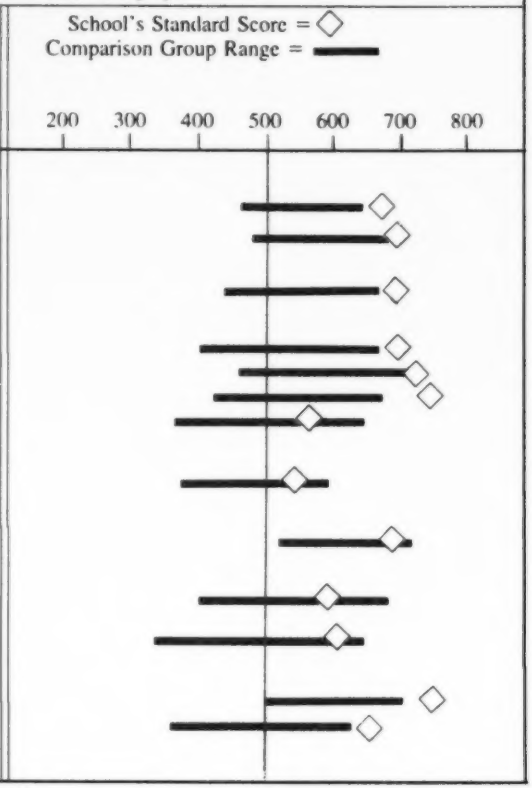
Many school districts and state agencies have found ways to communicate relatively detailed summary profiles of performance.[4] Compared to the practice of reporting only simple average scores on a standardized test, more elaborate profiles can give information on many variables (i.e., achievement, participation, school climate) according to student groups, departments, schools, and districts, and for different time periods.

In St. Paul, Minn., a profile is developed for each school using data about student achievement on standardized tests, school climate, attendance, course-taking, library usage, and student mobility. A section for each school portrays key physical and programmatic features, including special activities, programs, awards, distinctions, and the like. (St. Paul Public Schools, 360 Colbourne St., St. Paul, Minn. 55102.)

[4]Educational Research Service in Arlington, Va., as part of its service in providing management information resources, collects and makes available school district profiles and other research reports. Cooley and Bickel (1986) in *Decision Oriented Education Research* write about the use of school district-based research for administration and program improvement. Other districts that show creative reporting of multiple indicators and disaggregated test data include Detroit, Cincinnati, Fairfax County in Virginia, Los Angeles, Miami-Dade, Milwaukee, Pittsburgh, and Portland, Oreg.

Figure 3.2
Excerpt from a State High School Performance Report

Course Enrollments 1985-86	1985-86 School Standard Score	Comparison Group Standard Score Range
Mathematics		
3 or more years	679	463-653
Advanced mathematics	688	484-688
English		
4 or more years	620	438-646
Science		
3 or more years	671	430-648
Chemistry	694	452-690
Physics	721	440-668
Advanced science	528	378-626
History/Social Science		
4 or more years	522	404-610
Foreign Language		
3 or more years	688	507-702
Fine Arts		
1 year art/music/drama/dance	566	421-673
Units required for graduation	577	354-614
University of California Requirements		
Enrollments in a-f courses	709	495-676
Graduates completing a-f courses	646	386-633



Source: California State Department of Education, 721 Capital Mall, Sacramento, Calif. 95814.

These profiles use multiple indicators to create a more complete organizational portrait of each school. School mobility data and the disaggregation of test scores by income groups permit fairer comparisons across schools. The enrollment data provide information about course-taking patterns. Results from "self-observation" scales give a general measure of students' social attitudes related to the school and peers. The measure of library usage could be an interesting participation indicator.

The California Department of Education issues performance reports for each high school. Four basic sets of quality indicators are used:

- Enrollments in selected academic courses
- Statewide test scores
- Dropout and attendance rates
- Performance of college-bound students on national tests (SAT, Advanced Placement, College Board Achievement tests).

The report, exceeding 30 pages per school, allows each school to view its performance on each of these indicators in relation to statewide norms; its own past performance and projects; and the performance of other schools that have student bodies of similar socioeconomic composition (identified as "Comparison Group" in the data tables). For ease of comparison, the main indicators have been transformed into standard scores ranging from 200 to 800 with a mean of 500 for the state.

An example of information about course enrollments is given in Figure 3.2. Note that this school (Epic Senior High) generally places near the top of its comparison group and far above the state average in enrollment in academic subjects.[5] It falls closer to the average of the comparison group in history/social science and advanced science courses. Similar tables are provided for the other indicators, along with a three-year summary that gives a longitudinal picture.

Summary

This chapter has offered both general guidelines and specific examples for assessing the school as a whole. Guidelines emphasized the development of standards that make comparisons among schools possible and that allow for disaggregation of data to examine specific groups of students and different domains of performance and participation.

---

[5]A school's score may fall slightly outside the comparison group range because the horizontal bar excludes scores from the top and bottom 10 percent of schools making up the comparison group. Thus, the low point of the range of scores is represented by the score of the school at the 10th percentile for the comparison group and the high point is represented by the score of the school at the 90th percentile. This procedure makes the display more representative of the main body of schools in the comparison group.

Longitudinal standards are needed primarily to determine the effect of the school on achievement and to assess objectives unique to the school. Between-school comparisons are useful for assessing relative progress on those goals that all schools are expected to serve and to identify schools that need help and those that might provide it. In selecting standards for either longitudinal or between-school comparison, it is important to consider a range of indicators that include student participation, curriculum-aligned exercises, goals important for specific groups of students, and diverse types of quantitative indicators.

Assessing overall school performance is both necessary and inevitable. There is a risk, however, that in working to make these judgments meaningful at the state, national, or international levels, assessors will choose standardized summary indicators that obscure distinctive assets and accomplishments of individual schools and districts. If undue weight is given to indicators chosen for universal comparisons, incentives may be created to inflate scores at the expense of efforts to teach for more authentic forms of academic achievement. Accomplishments that may not be efficiently measured in a standardized way can, however, be authentic academic achievements that inspire pride and improve morale in a school.

School assessment requires a balance between the need for indicators that are easily quantifiable at higher organizational levels for millions of students and thousands of schools, and the need for more complex criteria for authenticity in academic achievement that are difficult to standardize across all classes and schools. Chapter 4 describes how these ideas can be incorporated into a general plan for high school assessment.

CHAPTER 4

Implementing Assessment Programs

WITHOUT PRESUMING THAT all high schools could or should adopt any single model of assessment, this chapter offers one possible way of pulling this material together. It suggests how a high school might approach assessment in a way that celebrates authentic forms of achievement, helps teachers improve instruction, and offers appropriate information for judging the quality of school performance.

General Guidelines

The principles proposed and the examples described here can be emulated and adapted, but they may also stimulate entirely new ideas. In developing local plans, keep four main guidelines in mind.

1. *Community Input*. The assessment program must be responsive to the unique circumstances of the community and the school. At the earliest stages it must involve teachers, administrators, school board representatives, and community members interested in assessment planning. To gain political support and to concentrate resources and attention on the project, participation must be as broad and representative as possible.

If criteria for authenticity and the need for multiple indicators are taken seriously, the assessment process could have repercussions not only within main academic courses, but throughout the curriculum, in cocurricular activities, in counseling services, and perhaps even in approaches to discipline. In launching an assessment project it is important, therefore, to involve curriculum authorities, counselors, other education specialists beyond teachers, and interested parties from community agencies, business, and institutions of higher education.

2. *Teacher Commitment*. If new forms of assessment are to improve instruction, teachers must view assessment as useful to their teaching, rather than as a necessary nuisance (or evil) imposed from the outside. If teachers see assessment primarily as a device to monitor their competence or to satisfy parents, taxpayers, or public officials, with no beneficial effect on instruction and learning, the project could actually undermine good teaching. Accountability to the community cannot be ignored, but it must be pursued in ways that inspire teachers to take pride in their students' progress and the achievements of the school as a whole. The principal can play a key role by supporting procedures that assure teachers that the main function of improved assessment will be to help them to teach rather than to evaluate and regulate them.

3. *Full Discussion of Assessment Problems and Proposals*. Consensus must be reached as to whether there is a need for improved assessment in the school and, if so, what specific issues should be addressed. There is probably room for improvement in any school that relies largely on data from standardized tests of general achievement such as the SATs or on typical competency exams.

Undoubtedly, some staff members will be satisfied with current procedures; others may see problems but be reluctant to devote much time or effort to change. Some may be enthusiastic about tackling the problem, especially if adequate resources are available. The principal should devise an appropriate strategy for initiating discussions, such as a meeting with a small representative group of respected staff members, with members of a particular department, with people from outside agencies who might help to finance the project, or with an official body such as a faculty cabinet.

Special study groups might be formed to handle specific issues most relevant to the school. The discussions could include presentations by outside authorities, but also by school staff members who wish to argue a point of view or present some of their own approaches.

Initially the discussions should avoid preoccupation with logistical problems; it is always possible to inhibit change by noting that current routines will be difficult to modify. Instead, dialog should focus on developing a principled approach to the issues raised (e.g., what do we mean by "competence"), to consideration of concrete alternatives, and, most important, to current practices by teachers that may offer illustrations of authentic tests of discrete competencies and of exhibitions.

In discussing current practices or proposals for improvement, for example, participants could be asked to address the following issues:

- What are the purposes of this assessment (accountability, improvement, selection-allocation), and what practices will actually ensure that the information produced will achieve the stated purposes?
- What is considered evidence of academic achievement and why? How do the indicators meet criteria for authenticity in academic achievement? That is, how do they offer evidence of disciplined inquiry, integration of knowledge, and value beyond evaluation?
- How does the assessment, whether individual or school-level, contribute to

the use of multiple indicators? To what extent do the indicators document change rather than competence at one time? To what extent do the indicators permit fair comparison with other schools?

- In the case of school-level indicators, how might participation indicators be collected and organized? How could quantitative data be disaggregated and reported more usefully? What indicators of school quality and academic success would be most meaningful to people in the community? How might authentic achievement be celebrated more effectively through communal public events?

4. *Think Big, Start Small.* Initial discussions should probe the most fundamental issues of educational purpose and how to organize and assess learning. Some of the possibilities could entail major changes in current practice. Given the complexities of high schools, however, and especially their difficulties in achieving consensus on how to teach students from diverse backgrounds, it will probably be wise to start by attacking only a small part of the problem.

The first step may be a pilot project in one department, a major revision of an important school or district test, an interdisciplinary team that focuses on assessing the integration of knowledge, a task force on organizational indicators, or an assessment center that serves as a clearinghouse for sharing and generating new ideas. Starting small maximizes the possibility of success in the short run, but the main challenge for the principal is to keep the larger, long-term vision prominent enough so that the smaller projects really do become steps toward significantly improved assessment in the school.

A Proposal

After a basic philosophy has been established, a long-range plan for the school might be proposed. The following proposal calls for innovative assessment within existing courses; departmental assessment of school curriculum; a yearly public fair to display the school's achievements; and a special pregraduation exhibition of mastery.

Improvements Within Courses

Each teacher, sharing ideas with departmental colleagues, develops activities for more authentic assessment of individual students within existing courses. This could involve a more authentic multiple-choice test of discrete competencies, expansion of writing and speaking tasks, or use of exhibitions to conclude major units or the course itself. After seeking feedback and endorsement from departmental colleagues, teachers submit the successful activities to a central clearinghouse that makes examples of these practices available to all faculty members.

48

Departmental Assessment of School Curriculum

To balance the emphasis placed on standardized test scores and to celebrate more visibly those curriculum goals that may not be aligned with external tests, departments also develop their own tests of discrete competencies. The departmental curriculum tests are given to a sample of students who took the courses targeted and to individual students who volunteer because they want these scores on their record.[1] In addition to conventional items, the tests include exercises that assess more authentic achievement and therefore require more flexible testing conditions.

To convey a vivid sense of the quality of student work, it would be useful to keep samples of the discourse, products, and performances that students generate for the tests. These could be collected for student cohorts during their first year, at a midpoint, and at graduation, showing growth of individual students who are representative of the class. The records could include both samples of the original work and scores or judges' ratings based on standard criteria.

The curriculum will not permit between-school comparisons unless they are developed in conjunction with other schools. However, the school can collect and report evidence, grounded in its own baseline standards, to demonstrate what students know and can do and what progress they make on selected tasks over a certain amount of time. Results should be disaggregated to examine the performance of students from different socioeconomic and curriculum track groupings.

The school-based curriculum tests will provide indicators—both individual and schoolwide—to complement whatever external standardized tests are used. Such curriculum tests administered consistently over a period of time will give indicators of school progress or decline in selected content areas.

Annual Achievement Fair

At a yearly public school fair, students and faculty members submit evidence of some significant academic accomplishment. Written products, video and audiotapes, physical artifacts, computer demonstrations, and dramatic performances can be included. Some of the exhibitions can be entered for special validation and competitive judging by experts. If it is not feasible to require all students to participate, each student may contribute at least one entry during his or her high school career. The entire community is invited to

[1]In reporting results, scores of volunteer students should be computed separately to avoid biasing the representative sample.

attend. Teachers devote time within courses for students to prepare their exhibitions. Teachers also would receive released time to prepare exhibitions that demonstrate the teachers' own mastery in their respective fields.

During the fair, the school administration displays data about the school. The school report might include:

- Profiles of the current students and staff members, how they have changed in the past few years, and changes expected in the future.
- Major curriculum objectives and requirements, samples of students' schedules and programs of study, enrollment patterns for different groups of students, comparisons to the curriculum and enrollment patterns of similar schools.
- Overall accomplishments of students during the year, including data about participation and proficiency on curriculum tests, samples of student work, and scores on external exams. Information can be summarized for the entire school, but also disaggregated for each class, for socioeconomic groups, and for students in different curriculum tracks. Whenever possible, comparisons between similar schools can be made.
- Accomplishments of the seniors presented in longitudinal form, showing progress since freshman year, and comparing current seniors with the previous senior cohort and with current seniors in comparable schools.

The school report could be supplemented by oral and/or written reactions, from a visiting team familiar with other comparable schools. This might be presented in a formal ceremony.

Pre-Graduation Exhibitions

During the semester prior to anticipated graduation, each student may enroll in an exhibition of mastery and portfolio development course that constitutes a capstone and final graduation requirement. The two purposes of the course are to complete a project that demonstrates authentic mastery of a topic or problem requiring synthesis of what has been learned in several courses during high school, and to collect in a portfolio a variety of evidence about what has been accomplished during the high school years (e.g., participation in school and non-school academic activities, summaries of the student's yearly accomplishments at the fair, letters of recommendation, grades, test scores, etc.). The projects might be assessed by committees similar to those used in the ROPE program, and the portfolios might be validated by community authorities.

These four parts make a comprehensive program aimed toward more authentic assessment. As suggested earlier, however, and depending on the unique needs and resources of the school, one might begin with only one of the parts or even a piece of one. Whatever the scope of the project, a vigilant watch will be necessary to make sure that it is implemented in ways that offer useful information and that assist teachers.

Summary

Assessment can serve such purposes as accountability, improvement, and selection-allocation. Accountability and selection-allocation often receive most attention, but they should not be pursued at the expense of school improvement. Assessment should be directed toward both individuals and organizations.

Assessment should aim toward authentic forms of achievement that are distinguished primarily by disciplined inquiry (substantive and procedural knowledge, in-depth understanding, and moving beyond prior knowledge); integration of knowledge; and value beyond evaluation (production of discourse, things, and performances through collaboration and flexible use of time).

Standardized tests of general academic achievement usually do not provide information useful for improving individual or school performance, and the forms of achievement they assess usually fall short of most criteria for authenticity.

A number of schools in the United States and elsewhere have used assessment practices that measure authentic achievement more faithfully. These include tests of discrete competencies in writing, speaking, and problem solving; exhibitions that show proficiency of larger chunks of competence; and portfolios and profiles that offer multiple indicators of individual achievement.

Improved organizational assessment of high schools would use extensive data on student participation to complement information on achievement; it would disaggregate data to show the performance of special groups, especially of students from low and high income families; it would gather information that permitted comparison on multiple indicators from year to year, and between schools.

Implementing these ideas should proceed on a school-by-school basis. A comprehensive plan might include teachers working for more authentic assessment within existing courses, departmental assessment of curriculum, an annual achievement fair, and a pre-graduation exhibition.

These proposals alone, of course, will not persuade resistant staff members of the need for improved assessment, nor will they address a number of logistical obstacles that might be raised even by sympathetic readers. The ideas could, however, begin a dialog that helps high school principals and others interested in the problem respond to concerns for accountability in ways that encourage teachers to teach, students to learn, and schools to celebrate their success.

High schools are under increasing pressures to "produce." They face demands from the business community, higher education institutions, government agencies, professional experts, parents, and students. But what should they be producing? Some demands give no more guidance than slogans or clichés. Some pose contradictory expectations. Some stretch the school's

human and financial resources beyond reasonable limits. In spite of these difficulties, the accountability movement presses ahead with new district and state curriculum requirements and with an escalation of testing mandated from above.

What kind of achievement should high schools promote? How to achieve effective assessment is not simply a technical issue to be solved by experts. The more complicated issue of educational purposes must be faced. Because of the momentum built up by the accountability movement, some may worry that even asking the question might derail the train. We suspect, however, that many accountability trains are on the wrong track or heading in the wrong direction, and that, therefore, it would be wise to slow down to take stock of where they are taking us and where we want to go.

APPENDIX

Uses and Limitations of Standardized Tests

ACADEMIC ACHIEVEMENT CAN BE ASSESSED in many ways, but when the public and policymakers seek evidence of school quality, they usually look to standardized tests. These tests have a potent influence on education policy and on public perceptions of schools.[1] Newspapers rank schools according to the scores, and legislators readily advocate statewide uniform standardized testing programs.[2] Within school districts, standardized tests are often viewed as the only solid measure of school quality, and many school improvement programs use performance on standardized tests as the principal measure of success.[3]

Standardized test scores allow simple comparisons between students, schools, districts, states, and nations. They are easily administered, take little time away from instruction, and, with a long history of use by psychometricians and major institutions, they carry scientific credibility.

---

[1]See the special issue on linking testing and instruction in *Journal of Educational Measurement*, Summer 1983. Airasian and Madaus (p. 103) write, "Increasingly, standardized achievement tests are being used for a host of policy-oriented purposes: assessing educational equity; providing evidence on school and program effectiveness; allocating compensatory funds to districts; evaluating teacher effectiveness; accrediting school districts; classifying students for remediation; and certifying successful completion of high school or a given grade of elementary school."

[2]One survey based on a random sample of 2,000 Minnesotans showed widespread support of statewide standardized testing with published comparisons of schools (Craig and Samaranayaka, 1985).

[3]According to a 1976 General Accounting Office report, 90 percent of sampled respondents from local education agencies use standardized norm-referenced tests to assess effects of school programs (Herman and Yeh, 1980).

Several different types of tests can be standardized. The focus here is on standardized tests of *general achievement* and *ability*—the tests most widely used in secondary schools to measure students' verbal, numerical, and analytical abilities. Some criticisms of these tests may not apply to standardized tests of knowledge in specific subjects such as science, literature, foreign language, or history.

The following discussion addresses three main concerns:

- The difficulty of gaining useful information due to the way all standardized tests are constructed and scored
- Special problems of general achievement and ability tests due to their insensitivity to curriculum in specific subject areas and their lack of predictive relationship to more authentic forms of achievement
- The tendency of items in all standardized tests (even those of specific subject areas) to neglect the assessment of depth of understanding, integration of knowledge, and production of discourse.

What Standardized Test Scores Mean

Imagine that students A, B, and C take a standardized test of general verbal achievement. The three students score at the 90th, 70th, and 50th national percentiles, respectively. What do these scores mean?

Rankings

The scores show the three students' standing in the norming population, that is, in the large sample of students used in the test development process. This sample is usually selected to be representative of a national population at a given age, say 16. Thus, the scores of students A, B, and C tell us that student A did better than 90 percent of the sample of 16-year-old testtakers; B, better than 70 percent; and C better than 50 percent of the testtakers. Beyond these relative rankings, the scores provide little information, due in part to the way the tests are constructed.

Standardized tests are designed so the scores of any representative population of students will be normally (bell-shaped) distributed—that is, 68 percent of the scores will always fall between a certain score above and a certain score below the mean.[4] To achieve these properties, developers of standardized tests write and try out several questions on students. The questions

[4]For instance, the Stanford-Binet IQ test is constructed so that 100 (the mean score) is "average," and 68 percent of the population will fall between 85 and 115.

selected for the final test are those that about half the students get right. Questions that most students get right or wrong are not used, because if just about everyone gets a question right or wrong that question does not help to rank students from high to low.

An often overlooked result of this process of test construction is that it is impossible for many students to experience relative success on standardized tests. To achieve a normal curve, the developers deliberately choose certain items to ensure that at least half the students will always score below average. This process also produces rankings that are influenced very little under normal circumstances, particularly during the short term, by school learning. A student's percentile ranking remains relatively constant from year to year, particularly as he or she reaches high school age.

A standardized test score is an accurate measure of a student's test-taking ability *relative to the norming population*, but how should we interpret the difference between students who score at the 90th, 70th, and 50th percentiles?

Non-Uniform Intervals in the Scale

A standardized test scale is not like a scale on a ruler where the numbers correspond to something that is precisely quantifiable and where differences between numbers have an exact and constant meaning. The additional number of questions a student must get correct to move up a given increment in percentile (e.g., 10 percentile points) varies, depending on where that student stands in the percentile rankings.

It is possible that while the percentile difference between students A, B, and C is the same (20 points), the difference between students A and B in number of questions answered correctly could be greater than the difference between students B and C. What's more, the additional questions answered correctly by student A are likely to be more difficult than those answered correctly by the others due to the design of standardized tests. Thus, student B may be significantly closer to student C than to student A in the areas of knowledge and ability measured by the test.

This characteristic limits the usefulness of standardized tests for measuring changes in performance for individuals who score either very high or very low in the distribution. Individuals in both categories can make exceptional gains in achievement, but because they may occupy the same relative place in the larger distribution, these gains may go undisclosed by a standardized post-test.

Grade Equivalents

The properties of standardized test scales also have implications for interpreting Grade Equivalents (GEs).

GEs, properly understood, can make test scores more meaningful than percentiles. They are derived by including in the norming sample of test-takers, students from several grades above and several grades below the grade level for which the test is designed. For instance, a 10th grade reading test is given to a national sample of 8th, 9th, 10th, 11th, and 12th graders. The average score for each of these groups is computed; thus, a 10th grade student's score can be compared not only to 10th grade norms, but also to the average score of 11th or 12th graders on the test.

If a 10th grader has a GE of 12 on a reading test, that means the 10th grader scored the same as the average 12th grader. It does not mean, however, the 10th grader reads 12th grade material as well as the average 12th grader. Reading performance on 12th grade material is not a factor in creating GEs for 10th grade.[5] Having reviewed the basis for percentile ranks and grade equivalents, it is now time to consider more directly the kinds of competence that the scores represent.

What the Questions on Standardized Tests Measure

Standardized test scores are commonly viewed as indicators of knowledge, abilities, traits, or achievements, and as predictors of future achievement, but there is disagreement about the kinds of achievements measured. Some people assume that a student who outperforms peers on a standardized test knows more; can better understand articles or books; can write or debate better; and can make more reasoned decisions.

At the other extreme, some critics claim that no such competencies have been demonstrated, that standardized tests measure no more than the ability to take multiple-choice tests that do not necessarily assess meaningful achievements or abilities. The most accurate interpretation is somewhere between these views.

---

[5]GEs can be misleading when used to make judgments about changes over time. As Coleman and Karweit (1972) write, "A student who remained exactly the same number of GEs behind (i.e., whose GE was the same from one year to the next) could in fact be moving up in percentile position! For example, if he [a 6th grader] had been at the 16th percentile in verbal ability at grade 6, 1.5 years behind, and was still 1.5 years behind at the 12th grade, he would have had to rise to the 32nd percentile to do so, that is to say, of the 84 percent of the students who were above him at grade 6, he would have had to pass 16 percent in order to achieve this position [the same 1.5 years "below" GE]."

Test-Taking Abilities and School Grades

Standardized tests of general achievement and ability measure test-taking ability. Students are asked to recall small bits of information or numerical formulas from memory; recognize incorrect written grammar; choose a best response to solve word analogy or "brain-teaser" logic questions;[6] infer unstated premises; and remember items from short reading passages. Proficiency on these sorts of tasks, particularly under tight time constraints, seems to be an ability that differs from person to person and remains relatively stable over time.[7] Under normal circumstances, a student who performs well on one general ability or achievement test will perform similarly well on almost any test of this type.

These tests also have a fairly strong relationship to grade point average, a correlation of about .5 (on a scale from 0 to 1.0). This probably reflects the fact that many school tests involve multiple-choice questions similar to those on general achievement tests. However, grades and general achievement scores measure different things. Grades are based on multiple-choice test questions, class discussion, writing assignments, projects, attentiveness, and effort in specific subjects. When students' GPAs are compared to their test scores from the previous year, the correlation drops; most studies show the correlation between high school senior year test scores and first year college grades to be about .35 (Linn, 1982).[8]

Authentic Academic Achievement

Standardized tests are not designed to measure the forms of understanding and competence suggested by many of the criteria for authentic academic achievement presented in Chapter 1. Performance on standardized tests of general achievement is a poor predictor of performance on tasks that require

---

[6]For example, "I have 5 black socks and 4 blue socks in a drawer. How many socks do I have to take out of the drawer to make sure I have a pair of the same color?" (From a standardized test of cognitive skills).

[7]See Jencks et al., 1972; Whimbey, 1985.

[8]A .5 correlation means that there is a 2/3 probability that a student in the middle of the GPA distribution will score somewhere between the 20th and the 80th percentile on a standardized test of general achievement or ability (like the SAT); the chances are one in three that the student might score above the 80th percentile or below the 20th percentile. With a .35 correlation, the chances are only about one in three that a 90th percentile senior will place in the top fifth of the college freshman class on grade point average (two out of three for placing in the top half). These probabilities are computed from the table in the Appendix.

disciplined inquiry, knowledge integration, and discourse production on novel problems.

At Alverno College in Milwaukee, Wis., assessment of student progress in the liberal arts curriculum is based on nontraditional procedures. One of these, the Integrated Competence Seminar (ICS), provides an opportunity for students to demonstrate integration of the competencies developed at the midpoint of their college education.

After studying background information, students deliver speeches to persuade a decision-making board to accept their proposals. Next, they complete the "In-Basket" exercise[9], responding to letters, memos, minutes, and reports, that all await action. They solve problems, set priorities, analyze, organize, and make decisions on seven different items as if they were board members encountering these situations in their offices (Mentkowski and Doherty, 1982).

Off-campus professionals trained in the use of assessment instruments for the ICS serve as assessors. Two assessors observe and evaluate each student's performance using coding schemes with specified criteria.[10] Students' performance on these activities, controlling for family background, had very low correlations (from .03 to .16) with their scores on several different types of standardized tests of cognitive skills taken concurrently with the ICS. Thus, knowing a student's standardized test score gives very little information for predicting that student's likely performance on the ICS.

Another study presented 6,000 beginning graduate students with several different types of "raw" research findings (data from field studies and experiments). The students were asked to formulate and write hypotheses they believed would explain the findings. This task involved an "ill-structured" problem—that is, one in which the question to be answered is not clearly defined, where the kind of information needed for a solution is not initially apparent, where all the information needed is not immediately available, and where there is no clear criterion for testing a proposed solution and no clearly defined process for applying a criterion. Rather than simply retrieving an answer from memory or applying a well-learned algorithm, students must organize the problem for themselves, applying "higher order" skills.[11]

The quantity and quality of the students' hypotheses were judged and scored by experts. Then scores from the "hypotheses" test were compared to the students' most recent standardized academic ability test scores (from the Graduate Record Examination). Low correlations, ranging from .18 to .26, were found.

---

[9]See Chapter 2, Part One, for an example of this exercise.

[10]Inter-rater reliability coefficients were .75 for the In-Basket, .72 for the oral presentation, and .78 for the combined ICS score. These coefficients indicate relatively high levels of agreement between assessors.

[11]Ward, Fredericksen, and Carlson (1980) citing Simon (1973).

A second part of the study found that student performance on formulating hypotheses was a better predictor than the standardized test scores of students' later accomplishments in graduate school tasks such as doing original research, designing and building laboratory equipment, and writing or co-writing a research report (Frederiksen and Ward, 1978; Ward, Frederiksen, and Carlson, 1980).[12]

Other studies have compared writing proficiency to standardized test scores of verbal ability. In these assessments, a sample of writing is read, evaluated, and given a numerical score.[13] Correlations between verbal standardized test scores and scores on holistic assessments of writing generally fall between .4 and .6 (Tyler, 1986). This shows that there is a relationship between performance on the two types of measures, but it would be risky to assume that a person who scored well on a standardized test of verbal ability could also write well.

The studies indicate that standardized tests of general achievement are poor indicators of student proficiency in tasks that differ markedly from the types of questions on the tests. This should be no surprise, but it does create a problem for schools that aim toward the kinds of achievements described in Chapter 1, or for schools that wish to assess more specific curriculum goals.

School Program and Curriculum

Standardized tests of general achievement or ability measure much non-school learning and are not very sensitive over the short run to specific program effects—that is, to the effects of changes in instructional methods or learning materials.[14] Two related reasons account for this. First, these tests are not tailored to any particular school's curriculum. They are developed to respond to a general national market. Second, to achieve their purpose of maximum discrimination between students, the questions used are those that about half of students will not be able to answer correctly. Questions that most students get right are discarded in the test development process.

---

[12]What about relationships between standardized test performance and later life outcomes or accomplishments? For people with similar socioeconomic backgrounds and the same level of education, standardized test scores contribute almost nothing to predicting later life outcomes. Research has examined relationships between standardized test scores and earnings (Jencks et al., 1979), scientific and artistic accomplishments (Munday and Davis, 1974), job competence, and other outcomes. Such studies failed to find consistent positive relationships.

[13]See Chapter 2, Part One for examples of direct measures of writing.

[14]Drastic changes in teaching would eventually affect test score changes. Crash programs on specific domains of test content and coaching programs can produce "abnormal" gains in test scores. Bangert-Drowns, Kulik, and Kulik (1983) found that coaching could raise scores by 1/4 of a standard deviation.

Thus, many students are tested on content for which they are likely to have received little in-school preparation. Conversely, content that is widely and effectively taught by schools is not tested.[15] The extent to which a standardized test covers a particular school's curriculum, and the extent to which a particular student has been taught what he or she is being tested on, is difficult to ascertain.

This accentuates the effect that prior learning at school and in the family has on test performance (Anastasi, 1982). Tests of general achievement and ability measure school learning in reading, math, and other subjects, as well as test-taking experience;[16] learning stemming from the linguistic environment of the home, and the student's exposure to reading opportunities, numerical problem-solving opportunities, and other sources of information (movies, TV, museums, magazines, parental conversation and explanations, learning games); and finally, traits that appear to be present at birth (cognitive and perceptual processes) which relate to test performance.[17]

Of course, all approaches to assessment reflect non-school influences. The problem is that standardized tests of general achievement accentuate the influence of those verbal skills and perceptual capacities developed over several years. Compared to an in-class final exam, a districtwide criterion-referenced test, or the International Baccalaureate exams, the learning that standardized tests measure varies more systematically with students' socioeconomic, cultural, and family characteristics.

This presents problems when these test results are used to compare schools. Schools may differ in test performance for reasons having little to do with program quality. In addition, because of the relationship between test performance and students' socioeconomic status, a school serving lower-income students can have a markedly greater educational effect on its students than a school serving more affluent families, yet still show lower test scores.

---

[15]For further discussion of this problem see Madaus, Airasian, and Kellaghan (1980); and Resnick and Resnick (1985).

[16]"Test-wiseness" is a partly, if not entirely, learned skill that "artificially" boosts test score performance. See Sarnacki (1979).

[17]The distinction commonly made between ability and achievement tests is more apparent than real. Green, in his presidential address to the American Psychological Association, said, "tests of general verbal and numerical skills are usually called aptitude tests, which is unfortunate since the term 'aptitude' seems to suggest an inborn, unchangeable trait. Actually the tests assess developed abilities—skills acquired through years of training and practice with verbal and numerical material" (Green, 1978).

Summary

Standardized tests of general achievement are used in most school systems in the United States, and with pressures for state-level monitoring of school districts mounting, their use is growing. Because important decisions are made on the basis of these tests, users should be well-informed of their functions and limitations.

The main purpose of general standardized tests is to efficiently and reliably discriminate between students (i.e., to rank them), so that about half the population tested will be above and half below a mean. However, the significance of distances between different scores or percentile rankings is uncertain, because the scores do not represent uniform intervals and because the sum of items answered correctly does not reflect proficiency on specific intellectual or technical tasks encountered beyond the test.

Beyond test-taking ability on multiple-choice items, the kinds of knowledge and abilities measured by general achievement tests is unclear. There is a moderately high correlation between test performance and grade point average, but for several reasons, standardized general achievement tests are unlikely to assess authentic academic achievement.

First, the types of questions run counter to the criterion of "disciplined inquiry," which emphasizes depth of factual and conceptual knowledge in particular academic domains. A smattering of nationally representative multiple-choice questions cannot reveal the extent to which disciplined inquiry has occurred.

Second, the tests offer few, if any, opportunities to demonstrate comprehension of integrated forms of knowledge. Instead, the test items sample a broad range of isolated pieces of knowledge or superficial familiarity with diverse information.

Third, authenticity calls for demonstrating mastery that is meaningful beyond the instructional setting, typically through the production of discourse and artifacts in collaboration with others and within a flexible time frame. Standardized tests require no products or discourse (other than a "bubble" sheet and pencil), prohibit collaboration, and are taken under fixed, tight time constraints.

The general failure to meet criteria of authenticity is supported by studies in which standardized test scores show relatively low correlations with more direct measures of students' generating and organizing ideas and with skills in communication and analytical problem solving.

Finally, the tests are generally unresponsive to specific aspects of high school curriculum, and student scores are heavily influenced by family background.

In spite of these limitations, standardized tests of general achievement provide student rankings according to national norms. They help to predict performance in school and on similar tests in the near future. They therefore help to describe the relative achievement of a student or a school on certain

types of tasks and can facilitate educational placement decisions. They can be administered to large numbers of students with minimal inconvenience and scored at reasonably low cost.

Standardized tests of general achievement are often used and interpreted inappropriately.[18] But they should not be criticized for failing to measure what they were never designed to measure.

---

[18]See Gould (1981), *The Mismeasure of Man*, an important treatment of this subject.

References

Applebee, A. N. *Writing in the Secondary School: English and the Content Areas*. Urbana, Ill.: National Council of Teachers of English, 1981.

Applebee, A. N. *Contexts for Learning To Write: Studies of Secondary School Instruction*. Norwood, N.J.: Ablex Publishing Co., 1984.

Adler, M. *The Paideia Proposal: An Educational Manifesto*. New York: Macmillan Co., 1982.

Airasian, P. W., and Madaus, G. F. "Linking Testing and Instruction: Policy Issues." *Journal of Educational Measurement* 2(1983):103-18.

Anastasi, A. *Aptitude and Achievement Tests: The Curious Case of the Indestructible Strawperson*. Paper presented at Invited Symposia: State of the Art Series—Achievement Testing, at the meeting of the American Psychological Association, Washington, D.C., August, 1982.

Archbald, D. A., and Witte, J. F. *Metropolitan Milwaukee Specialty Schools and Programs Report*. Report to the Study Commission on the Quality of Education in the Metropolitan Milwaukee Public Schools. Madison, Wis.: Wisconsin Center for Education Research, 1985.

Bangert-Drowns, R. L., Kulik, J. A.; and Kulik, C. C. "Effects of Coaching Programs on Achievement Test Performance." *Review of Educational Research* 4(1983):571-85.

Bloom, B. S.; Madaus, G.F.; and Hastings, J.T. *Evaluation To Improve Learning*. New York: McGraw-Hill, 1981.

Boyer, E. L. *High School: A Report on Secondary Education in America*. New York: Harper and Row, 1983.

Burgess, T., and Adams, E. *Records of Achievement at 16*. Windsor, Berkshire, SL4IDF, England: NFER-Nelson, 1985.

_____. *Outcomes of Education*. London: Macmillan & Co., 1980.

Burstein, L.; Baker, E. L.; and Aschbacher, P., with Keesling, J. W. *Using State Test Data for National Indicators of Education Quality: A Feasibility Study* (Final Report). Los Angeles, Calif.: University of California Center for the Study of Evaluation, 1986.

Coleman, J.S., and Karweit, N.L. *Information Systems and Performance Measures in Schools*. Englewood Cliffs, N.J.: Educational Technology Publications, 1972.

Committee for the Assessment of Experiential Learning. *A Guide for Assessing Prior Experience Through Portfolios* (Working Paper No. 6). Princeton, N.J.: Educational Testing Service, 1975.

Cooley, W. W., and Bickel, W. E. *Decision-Oriented Educational Research*. Boston, Mass.: Kluwer-Nijhoff Publishing, 1986.

Craig, W., and Samaranayaka, K. *1985 Minnesota Citizen Opinions on Public Education and Educational Policies*. Minneapolis, Minn.: Center for Urban and Regional Affairs, University of Minnesota, 1985.

Fielding, M., and Fiasca, M. *Personal and Social Issues in Science: Lessons from the Classroom. A Handbook for Secondary Science Teachers.* Monmouth, Oreg.: Teaching Research Division, Oregon State System of Higher Education, in press.

Forrest, A. *A Student Handbook on Preparing a Portfolio for the Assessment of Prior Experiential Learning.* CAEL Working Paper No.7. Princeton, N.J.: ETS, 1975.

Forrest, A.; and Steele, J. M. *Defining and Measuring General Education Knowledge and Skills* (Tech. Rep. No. 1976-81). Iowa City, Iowa: College Outcome Measures Project, The American College Testing Program, 1982.

Frederiksen, N. "The Real Test Bias: Influences of Testing on Teaching and Learning." *American Psychologist* 39(1984):193-202.

_____. "How Can the Higher-Order Skills be Measured?" A paper presented at a meeting of the Panel on Indicators of Mathematics and Science Education, National Research Council, Washington, D.C., June 1985.

Frederiksen, N.; and Ward, W. C. "Measures for the Study of Creativity in Scientific Problem Solving." *Applied Psychological Measurement* 2(1978):1-24.

Gardner, E. "Some Aspects of the Use and Misuse of Standardized Aptitude and Achievement Tests." In *Ability Testing: Uses, Consequences, and Controversies. Part II*, edited by A.K. Wigdor and R. Garner. Committee on Ability Testing. Washington, D.C.: National Academy Press, 1982.

Gardner, H. *Frames of Mind: The Theory of Multiple Intelligences.* New York: Basic Books, 1983.

Gibbons, M. "Walkabout: Searching for the Right Passage from Childhood and School." *Phi Delta Kappan* 9(1974):596-602.

_____. *The New Secondary Education: A Phi Delta Kappa Task Force Report.* Bloomington, Ind.: Phi Delta Kappa, 1976.

_____. "Walkabout Ten Years Later: Searching for a Renewed Vision of Education." *Phi Delta Kappan* 9(1984):591-600.

Goodlad, J. I. *A Place Called School: Prospects for the Future.* New York: McGraw-Hill, 1984.

Gorman, T. *The Framework for the Assessment of Language.* NFER-NELSON, Darville House, 2 Oxford Road East, Windsor, Berkshire SL41DF, England, 1986.

Gorman, T.; White, J.; and Brooks, G. *Language Performance in Schools.* Report on the 1982 secondary survey from the Language Monitoring team at the National Foundation for Educational Research in England and Wales to the Department of Education and Science, the Welsh Office and the Department of Education for Northern Ireland, 1982.

Gould, S. *The Mismeasure of Man.* New York: W. W. Norton, 1981.

Green, B. F., Jr. "In Defense of Measurement." *American Psychologist* 33 (1978):664-70.

Greenberg, K. L.; Wiener, H. S.; and Donovan, R. A., eds. *Writing Assessment: Issues and Strategies*. White Plains, N.Y.: Longman, Inc., 1986.

Herman, J.; and Yeh, J. "Test Use: A Review of the Issues." In *Educational Testing and Evaluation*, edited by E. Baker and E. Quellmalz. Beverly Hills, Calif.: Sage, 1980.

Heyns, B. *Summer Learning and the Effects of Schooling*. New York: Academic Press, 1978.

Hirsch, E. D. *Cultural Literacy: What Every American Needs To Know*. Boston: Houghton Mifflin, 1987.

Hogan, T. P., and Mishler, C. *Relationships Among Measures of Writing Skill*. Green Bay, Wis.: University of Wisconsin-Green Bay, 1981.

Jencks, C.; Bartlett, S.; Corcoran, M.; Crouse, J.; Eaglesfield, D.; Jackson, G.; McClelland, K.; Mueser, P.; Olneck, M.; Schwartz, J.; Ward, S.; and Williams, J. *Who Gets Ahead?* New York: Basic Books, Inc., 1979.

Jencks, C.; Smith, M.; Acland, H.; Bane, M. J.; Cohen, D.; Gintis, H.; Heyns, B.; and Michelson, S. *Inequality: A Reassessment of the Effect of Family and Schooling on America*. New York: Harper and Row, 1972.

Klein, K., ed. *Phi Delta Kappa Hot Topics Series: Student Competency Testing*. Bloomington, Ind.: Phi Delta Kappa, Center on Evaluation, Development and Research, 1983-84.

Linn, R. L. "Ability Testing: Individual Differences, Prediction, and Differential Prediction." In *Testing: Uses, Consequences, and Controversies, Part II*, edited by A. K. Wigdor and W. R. Garner. Washington, D.C.: National Academy Press, 1982.

Loacker, G.; Cromwell, L.; Fey, J.; and Rutherford, D. *Analysis and Communication at Alverno: An Approach to Critical Thinking*. Milwaukee, Wis.: Alverno College, 1984.

Maclure, M., and Hargreaves, M. *Speaking and Listening: Assessment at Age 11*. Published by NFER-NELSON, Darville House, 2 Oxford Road East, Windsor, Berkshire SL41DF, England, 1986.

Madaus, G. F.; Airasian, P. W.; Kellaghan, T. *School Effectiveness: A Reassessment of the Evidence*. New York: McGraw-Hill, 1980.

Mentkowski, M., and Doherty, A. *Validating Assessment Techniques in an Outcome-Centered Liberal Arts Curriculum: Integrated Competence Seminar*. Milwaukee, Wis.: Alverno College, 1982.

Munday, L., and Davis, J. *Varieties of Accomplishment After College: Perspectives on the Meaning of Academic Talent*. Iowa City, Iowa: American College Testing Research Report No. 62, 1974.

Murnane, R. J., and Raizen, S. A. *Improving Indicators of the Quality of Science and Math Education in Grades K - 12*. Washington, D.C.: National Academy of Sciences, 1988.

NAEP. *Learning by Doing: A Manual for Teaching and Assessing Higher-Order Thinking in Science and Mathematics*. The NAEP study was supported by the National Science Foundation through a grant to the Center for Statistics, Office for Educational Research and Improvement, U.S. Department of Education, 1987.

Newmann, F. M. *The Assessment of Discourse in Social Studies*. Paper commissioned by the Study Group on the National Assessment of Student Achievement. Madison, Wis.: National Center on Effective Secondary Schools, 1988. (1986 draft available from ERIC Document Reproduction Service No. ED 279 869.)

NPR. *Standardized Testing*, Parts 1-6, Program Nos. 280-285. Washington, D.C.: National Public Radio and the Institute for Educational Leadership, 1980.

Powell, A. G.; Farrar, E.; and Cohen, D. K. *The Shopping Mall High School*. Boston: Houghton Mifflin, 1985.

PROPEL. The Arts PROPEL Project, Harvard Project Zero, 326 Longfellow Hall, Cambridge, Mass, The Harvard Graduate School of Education, 1987.

Raizen, S. A., and Jones, L. V., eds. *Indicators of Pre-College Education in Science and Mathematics: A Preliminary Review*. Washington, D.C.: National Academy Press, 1985.

Resnick, D. P., and Resnick, L. B. "Standards, Curriculum, and Performance: A Historical and Comparative Perspective." *Educational Researcher* 4(1985):5-21.

Roberts, A. D., and Cawelti, G. *Redefining General Education in the American High School*. Washington, D.C.: Association for Supervision and Curriculum Development, 1984.

Royster, E.; Baltzell, D.; and Simmons, F. *Study of the Emergency School Aid Act Magnet School Program*. Washington, D.C.: USOE, 1979.

Sarnacki, R. E. "An Examination of Test-Wiseness in the Cognitive Test Domain." *Review of Educational Research* 2(1979):252-79.

Simon, H. A. "The Structure of Ill-Structured Problems." *Artificial Intelligence* 4(1973):181-201.

Sizer, T. S. *Horace's Compromise: The Dilemma of the American High School*. Boston: Houghton Mifflin, 1984.

Tyler, R. W. "Changing Concepts of Educational Evaluation." *International Journal of Educational Research* 1(1986):1-113.

Tyler, R. W., and White, S. H. *Testing, Teaching, and Learning*. Washington, D.C.: National Institute of Education, 1979.

Tyrrell, B., and Adams, E. *Records of Achievement at 16*. Windsor, Berkshire, SL4IDF, England: NFER-NELSON, 1985.

Ward, W. C.; Frederiksen, N.; and Carlson, S. B. "Construct Validity of Free-Response and Machine-Scorable Forms of a Test." *Journal of Educational Measurement*, Spring 1980, pp. 11-29.

Whimby, A. "You Don't Need a Special 'Reasoning' Test To Implement and Evaluate Reasoning Training." *Educational Leadership* 2(1985):37-39.

Wigdor, A. K., and Garner, W. R. *Ability Testing: Uses, Consequences, and Controversies, Part I*. Washington, D.C.: National Academy Press, 1972.

Williams, P.A., *Standardizing School Dropout Measures*, Madison, Wisc.: Center for Policy Research in Education, 1987.